WHEN POLICY IMPROVEMENT BREAKS FUTURE LEARNING: SELF-VALIDATING POLICY OPTIMIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Online reinforcement learning can invalidate its own next update. In continuing tasks with unbounded states, an approximate policy update can make the induced on-policy process heavy-tailed or poorly mixed. Subsequent value estimates and policy gradients may then become unreliable, even when the initial policy is stable and a stable optimum exists. To address this failure, we introduce a graded certificate for policy updates that identifies whether a raw policy preserves the statistical conditions needed by subsequent learning. We then develop *self-validating policy optimization* (SVPO), an average-cost actor–critic method that couples policy improvement to the validity of its own future updates. By retaining the physical task cost and preserving the certificate in the learned policy itself, SVPO keeps task performance aligned with the original objective and makes that policy valid for continued data collection and learning. The same construction yields Tilt, a certificate-preserving policy correction for existing policy optimizers. We prove that the certificate ensures that SVPO preserves the moment and mixing guarantees needed by future learning uniformly along its adaptive policy sequence. This establishes almost-sure convergence of the coupled actor–critic recursion to projected stationarity under standard assumptions. Experiments on switched LTI, input-saturated LQR, and unbounded queueing systems show that certificate-margin loss precedes downstream degradation, while SVPO maintains positive raw-policy margins with favorable physical task performance.

1 INTRODUCTION

Online reinforcement learning (RL) changes both control behavior and the Markov chain that generates data for the next update. This self-referential property is particularly important in continuing control problems with unbounded state spaces where states often represent cumulative quantities such as queue lengths, inventories, or energy imbalances (Neely, 2010; Dai & Gluzman, 2022). Average-cost evaluation and stochastic approximation rely on the policy-induced chain having a stable long-run distribution, controlled moments, and sufficient mixing (Tsitsiklis & Van Roy, 1999; Borkar & Meyn, 2000; Saxena et al., 2023; Ganesh et al., 2025). An approximate policy update can destroy these properties, even when learning starts from a stable policy and a stable optimum exists. The resulting state tails can amplify temporal-difference (TD) errors and learning noise, making subsequent value estimates and actor updates unreliable. We call this training-path failure *certificate collapse*.

Stability-aware methods intervene at several points in this feedback loop. Constrained and Lyapunov-based policy optimization enforce auxiliary-cost feasibility, including through policy-update constraints (Tessler et al., 2019; Achiam et al., 2017; Chow et al., 2018). Stability-oriented cost shaping modifies the training cost to favor stable behavior (Pavse et al., 2024), while action shields and barrier controllers restrict or modify actions at execution time (Alshiekh et al., 2018; Bastani et al., 2021; Ames et al., 2016; Cheng et al., 2019). These methods target constraint satisfaction or stable execution. Continued learning also requires regularity of the Markov chain induced by the raw policy used for sampling. We ask how approximate policy improvement can preserve this regularity at every update.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Our solution is to make learning validity an invariant of policy optimization. We formalize this invariant through a graded self-validation certificate. The required certificate grade yields the regularity needed for subsequent learning, including recurrence, moment and tail control, mixing, Poisson regularity, and noise bounds (Theorem 1, Corollary 2, Lemma 1 and Lemma 2). We then introduce self-validating policy optimization (SVPO), an online average-cost actor–critic that builds certificate preservation directly into each raw-policy update. Its positive certificate channel preserves the exponential certificate throughout learning (Theorem 2).

SVPO retains the physical task cost in its TD recursion and requires no execution-time filter. Its per-iterate certificate gives uniform control of on-policy Markovian noise and almost-sure convergence of the critic residual to zero and the actor to projected stationarity (Theorem 3). Tilt transfers the same mechanism to a candidate policy and gives the unique minimum-KL correction for its achieved certificate load.

Our experiments connect raw-policy certificate behavior to downstream learning. Switched LTI shows margin loss preceding learning deterioration. In input-saturated LQR, SVPO attains the lowest cumulative cost among the displayed certificate-preserving methods. In unbounded queueing, Tilt improves PPO's margin at similar cost, while SVPO reduces cost and tail occupancy.

This work makes three contributions.

1. **Learning-valid policy improvement.** A graded certificate characterizes a class of policies for subsequent learning regularity.
2. **Self-validating optimization.** SVPO preserves the certificate throughout online actor–critic learning with almost-sure convergence guarantees; Tilt provides the minimum-KL correction under its certificate-load constraint.
3. **Mechanism and experiments.** Across switched LTI, input-saturated LQR, and an unbounded queueing system, SVPO maintains positive raw-policy margins with favorable task and tail behavior, while Tilt improves PPO's margin with little change in task cost.

2 Preliminaries and Problem Formulation

2.1 General MDP with Average Cost

We consider a continuing Markov decision process $\mathcal{M} = (\mathcal{X}, \mathcal{A}, \mathsf{P}, c)$ on a Polish state space $\mathcal{X}$ and action space $\mathcal{A}$. The transition kernel and stage cost are

$$\mathsf{P} : \sigma(\mathcal{X}) \times \mathcal{X} \times \mathcal{A} \to [0,1], \qquad c : \mathcal{X} \times \mathcal{A} \to \mathbb{R}.$$

For a measurable function $f : \mathcal{X} \to \mathbb{R}$, write $\mathsf{P}f(x,a) = \int_{\mathcal{X}} f(x')\mathsf{P}(dx' \mid x,a)$. For a stationary policy π and any measurable set $B \in \sigma(\mathcal{X})$, let $\mathsf{P}_\pi(B \mid x) = \int_{\mathcal{A}} \mathsf{P}(B \mid x,a)\pi(da \mid x)$. The objective is the long-run average cost

$$J^\pi(x) := \limsup_{T\to\infty} \frac{1}{T} \mathsf{E}_{\mathsf{P}_\pi}\left[\sum\nolimits_{t=0}^{T-1} c(x_t,a_t) \,\middle|\, x_0 = x \right].$$

Average-cost evaluation uses a relative value satisfying a Poisson equation. Its existence and the control of trajectory-dependent learning noise depend on the recurrence, moments, and mixing of P_π.

We use a measurable Lyapunov function $L : \mathcal{X} \to [1,\infty)$. The task cost and features satisfy an envelope $C(1 + L^{q/k})$ in Assumption 5, with integers $0 < q < k$ and order gap $\Delta = k - q > 0$. The measurable dissipativity feature $\phi_H : \mathcal{X} \times \mathcal{A} \to [0,\infty)$ captures action-dependent growth; comparable features yield the same certificate grades (Proposition 1). Define $\mathcal{D} : \mathcal{X} \to [0,\infty)$ as a state-only drift budget. A fixed probability measure ν on $\mathcal{A}$ is used as the base measure.

Assumption 1 (Lyapunov dissipativity and stabilizability)**.** *There exist an L-bounded petite set $\mathcal{C}$, constants $b_0, b_1 \geq 0$ and $c_0 \in (0,1)$, and a stationary reference policy Π such that*

$$\mathsf{P}L(x,a) - L(x) \leq \phi_H(x,a) - \mathcal{D}(x) + b_1 \mathbb{I}_{\mathcal{C}}(x), \tag{1a}$$

$$\int_{\mathcal{A}} \phi_H(x,a)\Pi(da \mid x) - \mathcal{D}(x) \leq -c_0 L(x) + b_0 \mathbb{I}_{\mathcal{C}}(x). \tag{1b}$$

Averaging equation 1a under Π and using equation 1b gives geometric drift for the reference policy. Together with the standard petite-set and irreducibility conditions in Appendix A, the Foster–Lyapunov criterion of Meyn & Tweedie (2012) therefore ensures that the task is stabilizable. We ask which online actor–critic iterates preserve the resulting regularity.

2.2 Certificate-Preserving Policy Improvement

We now formulate the central question of this paper: *what makes a policy improvement step self-validating?* Precisely, what condition ensures that an online policy update preserves the Markov-chain stability conditions needed for its own future evaluation and improvement?

Equation 1a isolates the certificate quantity controlling the policy-dependent term in the drift bound. Define the excess dissipativity load by

$$s(x,a) := \phi_H(x,a) - u^*(x), \tag{2}$$

where $u^*(x) := \min_{a\in\mathcal{A}} \phi_H(x,a)$. The load $s(x,a) \geq 0$ measures the additional leading-order drift associated with choosing a instead of an action minimizing $\phi_H(x,\cdot)$. Using $u^*(x) \leq \int_{\mathcal{A}} \phi_H(x,a)\Pi(da \mid x)$ together with Assumption 1 gives

$$\mathsf{P}L(x,a) \leq (1-c_0)L(x) + s(x,a) + b'\mathbb{I}_{\mathcal{C}}(x), \tag{3}$$

where $b' = b_0 + b_1$. Within this bound, policy dependence enters through the conditional law of s. This separation yields the graded self-validation certificate below, which we call the certificate-preserving policy improvement (CPPI).

Definition 1 (Graded self-validation certificate)**.** For $\widetilde{m} \in \mathbb{N}$, a policy π satisfies CPPI($\widetilde{m}$) if for every $1 \leq j \leq \widetilde{m}$, there exists $C_j < \infty$ such that

$$\int_{\mathcal{A}} s(x,a)^j \pi(da \mid x) \leq C_j(1 + L(x)^{j(1-\Delta/k)}) \quad \forall x \in \mathcal{X}.$$

It satisfies CPPI(∞) if there exist $\gamma > 0$ and $0 \leq \theta_s < \infty$ such that for every $x \in \mathcal{X}$ and $s_0 \geq 0$,

$$\mathsf{P}_\pi(s(x,a) \geq s_0 \mid x) \leq \exp\left(\theta_s(1 + L(x)^{1-\Delta/k}) - \gamma s_0\right).$$

We call an online learning procedure *self-validating* if it optimizes the average-cost objective while every raw-policy iterate π_t satisfies Definition 1 at the grade required downstream. Each iterate then induces the Markov-chain regularity used by its subsequent policy evaluation and improvement.

3 Algorithm Design

We now develop Self-Validating Policy Optimization (SVPO), an online average-cost actor–critic, to keep every raw-policy iterate certified under the CPPI condition. The construction separates a low-order task component from the dissipativity feature ϕ_H, and transfers the same positive ϕ_H term from the critic to the learned policy.

3.1 Kernel-Based Function Approximation

For a finite dictionary $\{(x_j,a_j)\}_{j=1}^d$ and a positive semidefinite kernel K, define the task feature $\psi_K(x,a) := (K((x,a),(x_1,a_1)),\ldots,K((x,a),(x_d,a_d)))$. With $\theta = (\theta_{\text{task}},\theta_H)$ and $g_\theta(x,a) = \langle\theta_{\text{task}},\psi_K(x,a)\rangle$, the critic approximates a relative action value as

$$\widehat{Q}_\theta(x,a) = g_\theta(x,a) + \eta(\theta_H)\phi_H(x,a), \quad \eta(\theta_H) := \text{Softplus}(\theta_H) + \epsilon_0, \tag{4}$$

where $\text{Softplus}(\theta_H) = \log(1+\exp(\theta_H))$,the task-dependent component g_θ satisfies the growth condition in Assumption 5, and $\eta(\theta_H) \geq \epsilon_0 > 0$ is the learned coefficient of ϕ_H.

The critic induces the Gibbs raw policy

$$\pi_\theta(da \mid x) = \frac{\exp\left(-\widehat{Q}_\theta(x,a)/\tau\right)}{\int_{\mathcal{A}} \exp\left(-\widehat{Q}_\theta(x,a')/\tau\right)\nu(da')}\nu(da), \tag{5}$$

Algorithm 1 Self-Validating Policy Optimization

Input: ρ_0, θ_0, u_0, step sizes $\{\beta_t, \alpha_t\}$, $\tau, \epsilon_0 > 0$, and features ψ_K, ϕ_H

1: **for** $t = 0, 1, 2, \ldots$ **do**
2: Form π_{u_t} using the certificate channel of θ_t, with $\eta(\theta_H) \geq \epsilon_0$.
3: Sample $a_t \sim \pi_{u_t}(\cdot \mid x_t)$, observe c_t, x_{t+1}, and draw $\widetilde{a}_{t+1} \sim \pi_{u_t}(\cdot \mid x_{t+1})$.
4: $\Delta_t = c_t - \rho_t + \widehat{Q}_{\theta_t}(x_{t+1}, \widetilde{a}_{t+1}) - \widehat{Q}_{\theta_t}(x_t, a_t)$.
5: $\rho_{t+1} = \rho_t + \beta_t(c_t - \rho_t)$; $\theta_{t+1} = \Gamma_\Theta(\theta_t + \beta_t \Delta_t \zeta_{\theta_t}(x_t, a_t))$.
6: $u_{t+1} = \Gamma_U\Big(u_t - \alpha_t \widehat{\nabla}_u \mathcal{L}_{\mathrm{actor}}(u_t; \theta_t)\Big)$.
7: Use u_{t+1} and the certificate channel of θ_{t+1} in the next raw policy.
8: **end for**

with temperature $\tau > 0$. Equivalently, $\pi_\theta(\cdot \mid x)$ is the unique global policy-space minimizer of the statewise free energy

$$\mathcal{G}_x(\pi; \theta) := \int_{\mathcal{A}} \widehat{Q}_\theta(x, a)\pi(da \mid x) + \tau \mathrm{KL}(\pi(\cdot \mid x) \| \nu)\,. \tag{6}$$

This is the standard variational form of a regularized MDP update (Geist et al., 2019). The ϕ_H component therefore enters the raw policy before that policy generates data. Appendix C gives the derivation.

SVPO can also use a separately parameterized practical actor,

$$\pi_u(da \mid x) \propto \exp\Big(- [\, w_u(x, a) + \eta(\theta_H)\, \phi_H(x, a)\,]/\tau \Big)\, \nu(da). \tag{7}$$

The actor parameters appear in the lower-order task component w_u, while the ϕ_H term is shared with the current critic. At state x, the actor update follows the gradient of $\mathcal{G}_x(\pi_u; \theta)$ with respect to u. The statewise realization gap is

$$\epsilon_\pi(x; \theta, u) := \mathcal{G}_x(\pi_u; \theta) - \mathcal{G}_x(\pi_\theta; \theta) = \tau \mathrm{KL}(\pi_u(\cdot \mid x) \| \pi_\theta(\cdot \mid x)).$$

The shared channel cancels from the energy difference in their log-density ratio, giving the lower-order bound $\epsilon_\pi \leq 2 \sup_a |w_u - g_\theta|$ under Assumption 5.

3.2 Online Actor–Critic Recursion

Algorithm 1 updates the average-cost estimate and critic on the fast timescale β_t and the task actor on the slow timescale α_t. Here $\zeta_\theta = \nabla_\theta \widehat{Q}_\theta$, and Γ_Θ, Γ_U project onto bounded parameter sets. The invariant-state actor loss is $\mathcal{L}_{\mathrm{actor}}(u; \theta) = \mathsf{E}_{x \sim \mu_{\pi_u}}[\mathcal{G}_x(\pi_u; \theta)]$. For $a_t \sim \pi_{u_t}(\cdot \mid x_t)$, the actor uses the pre-update critic to form

$$\widehat{\nabla}_u \mathcal{L}_{\mathrm{actor}} = \left(\widehat{Q}_{\theta_t}(x_t, a_t) + \tau \log \frac{d\pi_{u_t}(\cdot \mid x_t)}{d\nu}(a_t) \right) \nabla_u \log \frac{d\pi_{u_t}(\cdot \mid x_t)}{d\nu}(a_t). \tag{8}$$

Conditioned on x_t and the current parameters, this estimator is unbiased for $\nabla_u \mathcal{G}_{x_t}(\pi_{u_t}; \theta_t)$ with the critic held fixed. Averaging over the invariant state distribution yields the actor semi-gradient used in the recursion.

For a candidate policy ϱ with nondegenerate exploration, Tilt reweights it as

$$\pi_\eta(da \mid x) \propto \exp(-\eta \phi_H(x, a)/\tau) \varrho(da \mid x).$$

The state-only component $u^*(x)$ cancels under normalization, leaving exponential control of the excess-load tail. Tilt therefore satisfies CPPI(∞) under the same nondegeneracy condition. Proposition 2 shows that Tilt is the unique minimum-KL correction under its expected excess-load constraint. The attained expected load decreases monotonically with η.

4 Stability and Convergence Guarantees

We first establish the regularity induced by CPPI, then verify CPPI(∞) for SVPO and use the resulting uniform bounds to analyze its two-timescale recursion.

Assumptions. We use the standard regularity framework for general-state-space average-cost RL and stochastic approximation (Meyn & Tweedie, 2012; Tsitsiklis & Van Roy, 1999; Borkar, 2008). Assumptions 1 and 2 give the system and higher-order moment conditions; Assumptions 3 and 4 give action coverage and irreducibility; Assumptions 5 and 7 control task and feature growth and smooth policy dependence; and Assumption 8 supplies the common local small-set bound used for uniform policy-family estimates. The appendix also states the standard compact-projection conditions, two-timescale step-size assumptions and local quantitative regularity for the critic-dependent SARSA recursion. Complete assumptions and proofs are provided in Appendices A, B, and C.

4.1 CPPI AND SYSTEM STABILITY

Theorem 1 (CPPI preserves stability regularity)**.** *Suppose Assumptions 1, 2, and 4 hold, and let π satisfy CPPI($\widetilde{m}$). For every $1 \le m \le \widetilde{m}$, there exist $\widetilde{\lambda}_{m,\pi} \in (0,1)$, $\widetilde{b}_{m,\pi} < \infty$, and an L-bounded petite set $\widetilde{\mathcal{C}}_\pi$ such that*

$$\mathsf{P}_\pi L^m(x) \le \widetilde{\lambda}_{m,\pi} L^m(x) + \widetilde{b}_{m,\pi} \mathbb{I}_{\widetilde{\mathcal{C}}_\pi}(x). \tag{9}$$

Consequently, the chain is positive Harris recurrent with a unique invariant measure μ_π. For any initial distribution with $\mathsf{E}[L^m(X_0)] < \infty$,

$$\sup_{t \ge 0} \mathsf{E}[L^m(x_t)] \le \max\left\{ \mathsf{E}[L^m(x_0)],\ \frac{\widetilde{b}_{m,\pi}}{1-\widetilde{\lambda}_{m,\pi}} \right\} < \infty, \quad \mu_\pi(L^m) \le \frac{\widetilde{b}_{m,\pi}}{1-\widetilde{\lambda}_{m,\pi}}. \tag{10}$$

Moreover, under Assumption 5, for $\mathcal{V}(x) = 1 + L^m(x)$, there exist $K_{m,\pi} < \infty$ and $\rho_{m,\pi} \in (0,1)$ such that, for every measurable $f : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$ with finite weighted norm $\|f\|_{\mathcal{V}} := \sup_{x,a} |f(x,a)|/\mathcal{V}(x)$,

$$|\mathsf{P}_\pi^n f(x,a) - \mu_\pi(f)| \le K_{m,\pi} \rho_{m,\pi}^n \|f\|_{\mathcal{V}} \mathcal{V}(x). \tag{11}$$

Here P_π denotes the state-action kernel $\mathsf{P}_\pi((x,a), dx'\, da') = \mathsf{P}(dx' \mid x,a)\pi(da' \mid x')$, with invariant measure $\mu_\pi(dx)\pi(da \mid x)$.

The order gap enters the tail scale through $q/k = 1 - \Delta/k$. Grade-one CPPI gives

$$\mathsf{P}_\pi L(x) \le (1-c_0)L(x) + A_\pi L(x)^{q/k} + B_\pi + b' \mathbb{I}_{\mathcal{C}}(x),$$

where $A_\pi, B_\pi > 0$ are finite coefficients for which the bound holds. Taking C_1 to be the grade-one CPPI constant, one may always choose $A_\pi = B_\pi = C_1$. We define the tail threshold scale

$$\Lambda_\pi := \max\left\{ 1, \left(\frac{4A_\pi}{c_0}\right)^{k/\Delta}, \frac{4B_\pi}{c_0} \right\}. \tag{12}$$

Corollary 1 (Tail control)**.** *For every $1 \le m \le \widetilde{m}$ and every tail multiplier $\gamma > 1$, the invariant measure μ_π satisfies*

$$\mu_\pi(L > \gamma \Lambda_\pi) \le \widetilde{b}_{m,\pi} \Big/ \Big((1-\widetilde{\lambda}_{m,\pi}) \gamma^m \Lambda_\pi^m \Big).$$

Grade one supplies the invariant distribution and a finite average task cost. Higher grades supply the moments used by the Poisson and stochastic-approximation arguments.

4.2 ALGORITHMIC GUARANTEES OF SVPO

Theorem 2 (SVPO is exponentially certificate-preserving)**.** *Under Assumptions 3 and 5, the exact Gibbs policy in equation 5 and practical policy in equation 7 satisfy CPPI(∞). Their exponential rate can be chosen as*

$$\gamma = \eta(\theta_H)/\tau \ge \epsilon_0/\tau.$$

Under Assumptions 1, 5, and 7, the grade-one comparison yields an explicit sufficient certificate strength for any prescribed tail scale. Specifically, for every $\Lambda_* > 1$, the condition $\eta(\theta_H) \ge \eta_{\min}(\Lambda_*)$ in equation 26 in Appendix C.5 guarantees $\Lambda_{\pi_u} \le \Lambda_*$. The resulting bound makes explicit how the required certificate strength depends on the order gap and the practical realization error. Thus, choosing $\epsilon_0 \ge \eta_{\min}(\Lambda_*)$ guarantees the same tail scale uniformly throughout learning. Under the conditions of Corollary 4 in Appendix C.6, a common positive certificate floor is necessary and sufficient for common CPPI(∞) constants.

Corollary 2 (Uniform drift, moments, and mixing)**.** *Suppose Assumptions 1, 2, 3, 4, 5, and 8 hold. Then every conclusion of Theorem 1 and Corollary 1 holds uniformly over the exact Gibbs and practical SVPO policy families for* $1 \le m \le \widetilde{m}$*. If* $\mathsf{E}[L^m(X_0)] < \infty$*, the adaptive SVPO path also satisfies* $\sup_{t\ge 0} \mathsf{E}[L^m(X_t)] < \infty$*.*

Proposition 2 characterizes the policy-space properties of the certificate mechanism. The exact Gibbs policy is the unique statewise free-energy minimizer, while Tilt is the unique minimum-KL correction of a supplied candidate policy under its achieved certificate-load constraint. The positive floor in equation 4 preserves CPPI(∞) throughout learning.

We next state the regularity used by the stochastic-approximation argument. Let $Z_t = (X_t, A_t)$, $z = (x, a)$, and $z' = (x', a')$. Use P_{π_u} for the state-action transition kernel $\mathsf{P}_{\pi_u}(z, dz') = \mathsf{P}(dx' \mid x, a)\pi_u(da' \mid x')$ and μ_{π_u} for its invariant measure. Here π_u includes the current shared coordinate θ_H in equation 7. Set $\Delta(\rho, \theta; z, z') := c(x, a) - \rho + \widehat{Q}_\theta(z') - \widehat{Q}_\theta(z)$. Fix a working grade $m \ge 2$ with $2m \le \widetilde{m}$ and write $\mathcal{V}(x) = 1 + L^m(x)$. Define the one-step update fields

$$H_c(z, z'; \rho, \theta) := \begin{pmatrix} c(x, a) - \rho \\ \Delta(\rho, \theta; z, z')\nabla_\theta \widehat{Q}_\theta(z) \end{pmatrix}, \quad H_a(z; u, \theta) := \hat{\nabla}_u \mathcal{L}_{\text{actor}}(u; \theta).$$

Their conditional fields and invariant means are

$$h_c(z; u, \rho, \theta) := \int H_c(z, z'; \rho, \theta)\mathsf{P}_{\pi_u}(z, dz'), \qquad h_a(z; u, \theta) := H_a(z; u, \theta),$$

$$\bar{h}_c(u, \rho, \theta) := \int h_c(z; u, \rho, \theta)\mu_{\pi_u}(dz), \qquad \bar{h}_a(u, \theta) := \int h_a(z; u, \theta)\mu_{\pi_u}(dz). \tag{13}$$

Lemma 1 (Regularity of the update mean fields)**.** *There exists a constant* $C_H < \infty$ *such that*

$$\|h_c(\cdot; u, \rho, \theta)\|_{\mathcal{V}} \le C_H(1 + |\rho|), \qquad \|h_a(\cdot; u, \theta)\|_{\mathcal{V}} \le C_H.$$

Moreover, the fields are Lipschitz in the same weighted norm:

$$\|h_c(\cdot; u, \rho, \theta) - h_c(\cdot; u', \rho', \theta')\|_{\mathcal{V}} \le C_H(1 + |\rho| + |\rho'|)\|(u, \rho, \theta) - (u', \rho', \theta')\|,$$
$$\|h_a(\cdot; u, \theta) - h_a(\cdot; u', \theta')\|_{\mathcal{V}} \le C_H\|(u, \theta) - (u', \theta')\|. \tag{14}$$

The invariant means $\bar{h}_c$ *and* $\bar{h}_a$ *satisfy the corresponding Lipschitz bounds.*

Lemma 2 (Poisson equation and Markov noise)**.** *For any* f *with* $\|f\|_{\mathcal{V}} < \infty$*, the Poisson equation* $\widehat{f} - \mathsf{P}_{\pi_u}\widehat{f} = f - \mu_{\pi_u}(f)$ *admits a centered solution* $\widehat{f}$ *satisfying* $\|\widehat{f}\|_{\mathcal{V}} \le C_{\mathcal{V}}\|f\|_{\mathcal{V}}$*. Moreover, as* $n \to \infty$*,* $\sum_{t=0}^n \beta_t(h_c(Z_t; u_t, \rho_t, \theta_t) - \bar{h}_c(u_t, \rho_t, \theta_t))$ *and* $\sum_{t=0}^n \alpha_t(h_a(Z_t; u_t, \theta_t) - \bar{h}_a(u_t, \theta_t))$ *converge almost surely.*

Theorem 3 (SVPO convergence)**.** *Under Assumptions 1–8, the iterates generated by Algorithm 1 satisfy* $\|\bar{h}_c(u_t, \rho_t, \theta_t)\| \to 0$ *almost surely. The exact Gibbs implementation satisfies the same limit, with its invariant mean field defined under* π_{θ_t}*. Let* $(\rho^*(u), \theta^*(u))$ *denote the unique critic equilibrium of practical SVPO for fixed* u*, after fixing the additive constant. If* $\bar{h}_a(u, \theta^*(u)) = \nabla F(u)$ *for some continuously differentiable function* F *on* U*, then Algorithm 1 also satisfies*

$$\|u_t - \Gamma_U[u_t - \bar{h}_a(u_t, \theta_t)]\| \to 0 \qquad \textit{a.s.}$$

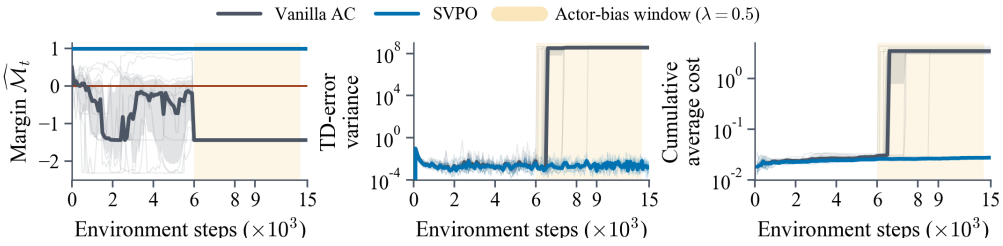
If the actor projection is inactive at the limit points, then $\|\nabla F(u_t)\| \to 0$ *almost surely.*

The exact realization uses average-cost SARSA updates. For practical SVPO, fixing the actor parameter u still leaves the behavior policy coupled to the critic through θ_H. The proof therefore first analyzes the average-cost and critic updates with u fixed, then the actor update on the slower timescale. The additional conditions in Appendix C.13 strengthen the convergence result to $J^{\pi_t} \to J^*$ almost surely (Corollary 5).

Corollary 3 (Comparator performance bound)**.** *Suppose Assumptions 1, 2, 3, 4, and 5 hold. Let* π *be a stationary comparator whose average-cost Poisson equation has a solution* Q^π *with* $\|Q^\pi - \widehat{Q}_{\theta_t}\|_{\mathcal{V}} < \infty$*, and suppose* $\pi(\cdot \mid x) \ll \nu$ *with* $\sup_x \mathrm{KL}(\pi(\cdot \mid x)\|\nu) < \infty$*. Then every practical SVPO iterate* $\pi_t = \pi_{u_t}$ *satisfies*

$$J^{\pi_t} - J^\pi \le \left(1 + \widetilde{b}_{m,\pi_t}/(1 - \widetilde{\lambda}_{m,\pi_t})\right)\left(2\|Q^\pi - \widehat{Q}_{\theta_t}\|_{\mathcal{V}} + \bar{\epsilon}_\pi\right) + \tau \sup_x \mathrm{KL}(\pi(\cdot \mid x)\|\nu). \tag{15}$$

For any stationary average-cost optimum satisfying the comparator conditions, setting $\pi = \pi^*$ gives the corresponding optimality-gap bound.



(a) Raw-policy margin (b) TD variability (c) Task performance

Figure 1: **Switched-LTI learning dynamics.** Panels show raw-policy margin, TD-error variance, and cumulative average task cost. Across ten matched seeds, thin curves show individual runs, thick curves medians, and bands interquartile ranges. Shading marks the controlled actor-bias window. The horizontal axis is linear through 9,000 steps and logarithmic thereafter; terminated Vanilla curves retain their final observation for visualization.

5 Experiments

We study three questions spanning mechanism, algorithm, and transfer. First, does certificate loss in approximate policy improvement precede degradation in subsequent learning? Second, can SVPO preserve the raw-policy certificate under continuous actions and critic approximation while retaining task performance, without execution-time correction? Third, can the CPPI-derived Tilt transfer certificate preservation to an existing learner, and does this mechanism remain effective in an unbounded scheduling problem? We study these questions using switched LTI, input-constrained LQR, and an MEC-motivated unbounded queueing system, respectively. Comparisons use matched objectives, initial task parameters, and random seeds, and certificates are evaluated on the learned raw policy.

At checkpoint t, we freeze the raw policy and evaluate

$$\widehat{\mathcal{M}}_t := \min_{x \in \mathcal{G}_{\text{tail}}} \frac{L(x) - \mathbb{E}_{a \sim \pi_t(\cdot|x), X^+ \sim \mathsf{P}(\cdot|x,a)}[L(X^+)]}{L(x)}.$$

This raw-policy margin provides a common empirical measure of whether the policy remains certified. We report it together with tail occupancy, TD-error variance, physical escape, and cumulative average task cost to assess sustained learning and task performance. Appendix D provides experimental protocols and extended results.

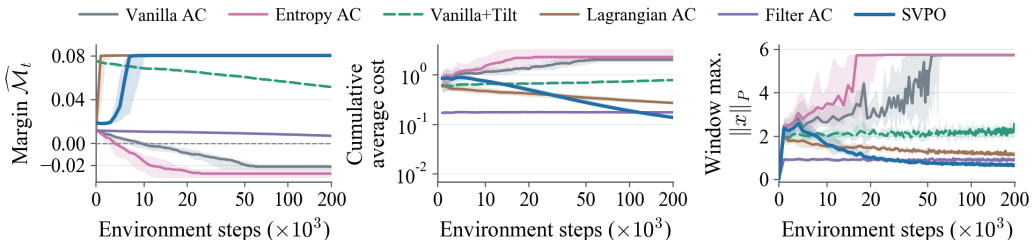
5.1 Switched LTI Systems

Switched LTI models mode-switching and gain-scheduled control, where a learner selects among closed-loop feedback modes. We use

$$X_{t+1} = f_{A_t} X_t + W_t,$$

with both contractive and expansive modes, a standard quadratic task cost, and a stable stationary policy. During learning, we temporarily introduce a lower-order actor bias that favors an expansive mode while leaving the dynamics, task cost, critic target, and Lyapunov function unchanged. This creates a controlled actor-realization error without changing which policies are stabilizing.

Figure 1 provides the mechanism audit. For the representative bias strength $\lambda = 0.5$, the bias is activated at step 6,000 and immediately drives Vanilla AC's already weakened margin to approximately -1.44. The rise in TD-error variance and cumulative task cost follows, with all ten seeds escaping between steps 6,086 and 8,576. SVPO preserves a positive margin and continues updating with controlled learning noise; its median cumulative average task cost is 0.02665 over 50,000 steps, versus 3.4684 for Vanilla AC up to escape. Across all three tested bias strengths, the margin and state-tail responses in Figure 5 in Appendix D show the same qualitative mechanism. This temporal ordering places certificate degradation before subsequent learning deterioration. Table 2 tests the effect of including the certificate in the sampling policy.



(a) Raw-policy margin (b) Task performance (c) State growth

Figure 2: **Input-saturated LQR.** Panels show raw-policy margin, cumulative average task cost, and the maximum state norm over the rolling window. Curves are medians with 10th–90th percentile bands over ten matched seeds. For Filter AC, margins use the raw policy; cost and state growth use executed actions. Terminated Vanilla AC and Entropy AC curves retain their final observation for visualization. The horizontal axis is linear through 10^4 steps and logarithmic thereafter.

5.2 Input-Saturated LQR

Input-saturated LQR represents actuator-limited continuous control, where a bounded physical input has a long-term effect on the state. We write

$$x_{t+1} = Ax_t + Bu_t + w_t, \qquad u_t \in \mathcal{U}.$$

All methods share an action-aware actor and a restricted task critic, creating a controlled representation limitation with unchanged dynamics and task objective. We compare Vanilla AC, Entropy AC, Vanilla+Tilt, Lagrangian, Filter AC, and SVPO variants; full details are in Appendix D.

Figure 2 shows that Vanilla+Tilt, Lagrangian AC, Filter AC, and SVPO maintain positive margins and complete training, whereas the raw-policy margins of Vanilla AC and Entropy AC fall below zero and escape in every seed. All methods are evaluated with the same physical quadratic cost but incorporate stability differently. Lagrangian AC adds a certificate penalty to its TD objective, whereas Filter AC corrects sampled actions using the known dynamics. SVPO keeps the physical task cost unchanged in its TD target and handles certificate preservation through a separate channel, keeping task-value learning directly aligned with the original control objective. Consistent with this distinction, SVPO attains the lowest median cumulative average physical cost without execution-time correction among the displayed certificate-preserving methods, achieving 0.1383 compared with 0.2713 for Lagrangian AC.

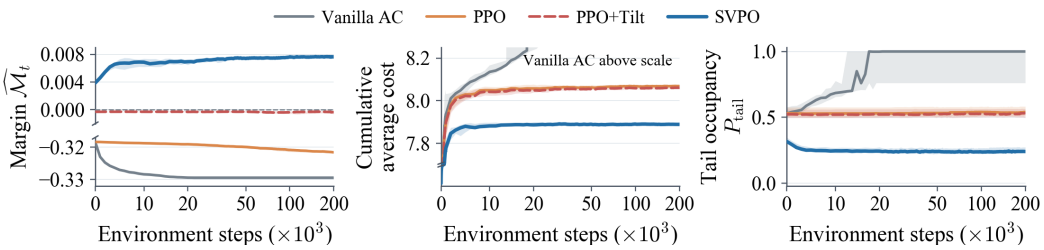
5.3 Unbounded Queueing Systems

Finally, we consider an unbounded queueing system with workload states and service-selection actions. Its dynamics are

$$q_{t+1} = F_{a_t} q_t + A_t^{\mathrm{arr}}.$$

The selected queue contracts while the unserved queue accumulates work. The system uses a non-negative energy–delay cost and admits a stable randomized scheduling policy.

The queueing case evaluates portability beyond the control settings. We compare Vanilla AC, PPO (Schulman et al., 2017), PPO+Tilt, and SVPO. Applying Tilt to PPO raises the final worst-case finite-grid raw-policy margin from -0.3216 to -0.0003, while yielding a slightly lower cumulative average task cost, with 8.0608 for PPO+Tilt versus 8.0689 for PPO. This result is consistent with Tilt's statewise minimum-KL construction under its achieved certificate-load constraint. Figure 3 shows that SVPO further increases the final margin to 0.0077 while reducing cumulative average task cost to 7.8891 and queue-tail occupancy to 0.2403, compared with 0.5317 for PPO+Tilt. Vanilla AC loses its certificate and reaches the physical threshold in all ten seeds, while PPO completes training without reaching the far-tail grid on which its worst-case margin violation is measured. Together with the two control cases, these results show that Tilt is portable across learning architectures and improves PPO's raw-policy margin. SVPO maintains positive margins throughout training while improving the reported task and tail metrics.



(a) Raw-policy margin (b) Task performance (c) Queue-tail occupancy

Figure 3: **Unbounded queueing system.** Panels show raw-policy margin, cumulative average task cost, and queue-tail occupancy. Curves are medians with 10th–90th percentile bands over ten matched seeds. Task cost is averaged from the first training step. Vanilla AC's margin and tail occupancy retain their final values after training stops. Horizontal axes are linear through 10^4 steps and logarithmic thereafter.

6 Related work

Average-reward policy-gradient and actor–critic methods rely on stationary distributions and differential values (Sutton et al., 1999; Konda & Tsitsiklis, 2003). TD and stochastic approximation analyses quantify the roles of mixing and Markovian noise in learning (Tsitsiklis & Van Roy, 1999; Suttle et al., 2023; Haque & Maguluri, 2025). Foster–Lyapunov theory provides recurrence, moment bounds, and ergodicity (Meyn & Tweedie, 2012), and exact average-cost policy iteration can propagate regularity through exact evaluation and improvement (Meyn, 1997; Hernández-Lerma & Lasserre, 1997). We instead study when this regularity is preserved under approximate online updates, where each updated policy generates the data used by the next learning step.

In **unbounded-state RL**, Shah et al. (2020) construct stable online policies using a sparse-sampling Monte Carlo oracle; Murthy et al. (2024) analyze natural policy gradient under a prescribed drift condition. Pavse et al. (2024) improve credit assignment through Lyapunov cost shaping and state transformations. Our approach directly constrains approximate raw-policy improvement to preserve the regularity required by subsequent learning. **Constrained policy optimization** and Lyapunov methods for constrained MDPs target cumulative-cost feasibility (Achiam et al., 2017; Chow et al., 2018). Lyapunov-based control and RL methods seek closed-loop stability (Berkenkamp et al., 2017; Chang et al., 2019; Han et al., 2020), while barrier controllers and shields restrict executed actions (Ames et al., 2016; Cheng et al., 2019; Alshiekh et al., 2018; Bastani et al., 2021). By contrast, our approach is imposed on the updated raw policy itself so that its induced chain retains the moment, mixing, and learning-noise bounds required by subsequent learning. **Entropy regularization and mirror descent** supply Gibbs policies and KL update geometry (Neu et al., 2017; Haarnoja et al., 2018; Geist et al., 2019; Grudzien et al., 2022). Tilt applies the same Gibbs/KL geometry to certificate preservation and gives the unique minimum-KL correction at its achieved certificate load. SVPO integrates certificate preservation into online average-cost actor–critic learning by using the same positive certificate coefficient in the critic and raw-policy update while retaining the physical task cost in its TD recursion. Under the stated assumptions, this structure preserves learning regularity throughout the updates and supports almost-sure convergence.

7 Conclusion

In continuing RL, policy improvement must preserve the conditions needed for future learning. Our graded self-validation certificate formalizes this requirement, and SVPO preserves it at each raw-policy update with an online actor–critic method. Tilt provides the unique minimum-KL correction at its achieved certificate load. The theory shows how this policy structure preserves learning regularity and supports the convergence analysis of SVPO. Experiments on switched LTI, input-saturated LQR, and unbounded queueing show that certificate preservation improves raw-policy learning regularity while maintaining and improving task performance.

AI USE STATEMENT

Generative AI tools were used to assist with language polishing, literature search and code development. All AI-assisted work was reviewed by the authors, who take full responsibility for the final content of the paper.

REPRODUCIBILITY STATEMENT

Appendices A–C provide the theoretical assumptions and proofs. Appendix D describes the experimental configurations, initialization, evaluation procedures, and extended results.

REFERENCES

Joshua Achiam, David Held, Aviv Tamar, and Pieter Abbeel. Constrained policy optimization. In *International conference on machine learning*, pp. 22–31. PMLR, 2017.

Mohammed Alshiekh, Roderick Bloem, Rüdiger Ehlers, Bettina Könighofer, Scott Niekum, and Ufuk Topcu. Safe reinforcement learning via shielding. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Aaron D Ames, Xiangru Xu, Jessy W Grizzle, and Paulo Tabuada. Control barrier function based quadratic programs for safety critical systems. *IEEE transactions on automatic control*, 62(8): 3861–3876, 2016.

Osbert Bastani, Shuo Li, and Anton Xu. Safe reinforcement learning via statistical model predictive shielding. In *Robotics: Science and Systems*, volume 17, pp. 10–15607, 2021.

Felix Berkenkamp, Matteo Turchetta, Angela Schoellig, and Andreas Krause. Safe model-based reinforcement learning with stability guarantees. *Advances in neural information processing systems*, 30, 2017.

Vivek S. Borkar. *Stochastic Approximation: A Dynamical Systems Viewpoint*. Cambridge University Press, 2008.

Vivek S Borkar and Sean P Meyn. The ode method for convergence of stochastic approximation and reinforcement learning. *SIAM Journal on Control and Optimization*, 38(2):447–469, 2000.

Ya-Chien Chang, Nima Roohi, and Sicun Gao. Neural lyapunov control. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Richard Cheng, Gábor Orosz, Richard M Murray, and Joel W Burdick. End-to-end safe reinforcement learning through barrier functions for safety-critical continuous control tasks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pp. 3387–3395, 2019.

Yinlam Chow, Ofir Nachum, Edgar Duenez-Guzman, and Mohammad Ghavamzadeh. A lyapunov-based approach to safe reinforcement learning. *Advances in neural information processing systems*, 31, 2018.

Jim G Dai and Mark Gluzman. Queueing network controls via deep reinforcement learning. *Stochastic Systems*, 12(1):30–67, 2022.

Swetha Ganesh, Washim Uddin Mondal, and Vaneet Aggarwal. A sharper global convergence analysis for average reward reinforcement learning via an actor-critic approach. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 18206–18227, 2025.

Matthieu Geist, Bruno Scherrer, and Olivier Pietquin. A theory of regularized markov decision processes. In *International conference on machine learning*, pp. 2160–2169. PMLR, 2019.

Jakub Grudzien, Christian A Schroeder De Witt, and Jakob Foerster. Mirror learning: A unifying framework of policy optimisation. In *International conference on machine learning*, pp. 7825–7844. PMLR, 2022.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International conference on machine learning*, pp. 1861–1870. PMLR, 2018.

Minghao Han, Lixian Zhang, Jun Wang, and Wei Pan. Actor-critic reinforcement learning for control with stability guarantee. *IEEE Robotics and Automation Letters*, 5(4):6217–6224, 2020.

Shaan Ul Haque and Siva Theja Maguluri. Stochastic approximation with unbounded markovian noise: A general-purpose theorem. In *Proceedings of the 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pp. 3718–3726, 2025.

Onésimo Hernández-Lerma and Jean B Lasserre. Policy iteration for average cost markov control processes on borel spaces. *Acta Applicandae Mathematica*, 47(2):125–154, 1997.

Vijay R Konda and John N Tsitsiklis. On actor-critic algorithms. *SIAM journal on Control and Optimization*, 42(4):1143–1166, 2003.

Francisco S Melo, Sean P Meyn, and M Isabel Ribeiro. An analysis of reinforcement learning with function approximation. In *Proceedings of the 25th international conference on Machine learning*, pp. 664–671, 2008.

Sean P. Meyn. The policy iteration algorithm for average reward markov decision processes with general state space. *IEEE Transactions on Automatic Control*, 42(12):1663–1680, 1997.

Sean P Meyn and Richard L Tweedie. *Markov Chains and Stochastic Stability*. Springer Science & Business Media, 2012.

Yashaswini Murthy, Isaac Grosof, Siva Theja Maguluri, and R Srikant. Performance of npg in countable state-space average-cost rl. *arXiv preprint arXiv:2405.20467*, 2024.

Michael J. Neely. *Stochastic Network Optimization with Application to Communication and Queueing Systems*, volume 3 of *Synthesis Lectures on Communication Networks*. Morgan & Claypool Publishers, 2010.

Gergely Neu, Anders Jonsson, and Vicenç Gómez. A unified view of entropy-regularized markov decision processes. *arXiv preprint arXiv:1705.07798*, 2017.

Brahma S Pavse, Matthew Zurek, Yudong Chen, Qiaomin Xie, and Josiah P. Hanna. Learning to stabilize online reinforcement learning in unbounded state spaces. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 40015–40033, 2024.

Naman Saxena, Subhojyoti Khastagir, Shishir Kolathaya, and Shalabh Bhatnagar. Off-policy average reward actor-critic with deterministic policy search. In *International Conference on Machine Learning*, pp. 30130–30203. PMLR, 2023.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Devavrat Shah, Qiaomin Xie, and Zhi Xu. Stable reinforcement learning with unbounded state space. In *Proceedings of the 2nd Conference on Learning for Dynamics and Control*, volume 120 of *Proceedings of Machine Learning Research*, pp. 581–581, 2020.

Wesley A Suttle, Amrit Bedi, Bhrij Patel, Brian M Sadler, Alec Koppel, and Dinesh Manocha. Beyond exponentially fast mixing in average-reward reinforcement learning via multi-level monte carlo actor-critic. In *International Conference on Machine Learning*, pp. 33240–33267. PMLR, 2023.

Richard S Sutton, David McAllester, Satinder Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. *Advances in neural information processing systems*, 12, 1999.

Chen Tessler, Daniel J. Mankowitz, and Shie Mannor. Reward constrained policy optimization. In *International Conference on Learning Representations*, 2019.

John N Tsitsiklis and Benjamin Van Roy. Average cost temporal-difference learning. *Automatica*, 35(11):1799–1808, 1999.

A TECHNICAL ASSUMPTIONS

The state and action spaces are Polish, $\mathsf{P}(\cdot \mid x, a)$ is a Borel transition kernel, and all policies below are stationary Markov kernels.

For a stationary policy π, let $c_\pi(x) = \int c(x,a)\pi(da \mid x)$. A relative value V^π solves

$$V^\pi(x) - \mathsf{P}_\pi V^\pi(x) = c_\pi(x) - J^\pi,$$

and $Q^\pi(x,a) = c(x,a) - J^\pi + \mathsf{P}V^\pi(x,a)$ whenever these quantities are well defined.

A.1 CONDITIONS FOR THE CPPI AND SVPO RESULTS

Assumption 2 (Higher-order Lyapunov regularity)**.** *There exists an integer $\widetilde{m} \geq 1$ such that, for each $1 \leq m \leq \widetilde{m}$ and $\varepsilon > 0$, there exists $R_\varepsilon < \infty$ such that*

$$\mathsf{P}L^m(x,a) \leq (1+\varepsilon)(\mathsf{P}L(x,a))^m$$

whenever $L(x) > R_\varepsilon$, uniformly over a. Moreover, $\sup_{L(x)\leq R_\varepsilon, a} \mathsf{P}L^m(x,a) < \infty$. Every set $\{x : L(x) \leq R\}$ is petite for each chain in Theorem 1.

Assumption 3 (Action coverage)**.** *The minimum $u^*(x)$ in equation 2 is attained and measurable in x. There exist constants $\xi_0 < \infty$ and $p_0 > 0$ such that*

$$\nu\{a : s(x,a) \leq \xi_0\} \geq p_0 \qquad \text{for every } x.$$

Assumption 4 (Irreducibility)**.** *The reference chain and every chain to which Theorem 1 is applied are ψ-irreducible and aperiodic.*

A.2 LEARNING ASSUMPTIONS

Assumption 5 (Task and certificate growth)**.** *There exist $C_{\mathrm{task}}, C_\phi < \infty$ such that, uniformly over the projected parameters,*

$$|g_\theta(x,a)| + |w_u(x,a)| \leq C_{\mathrm{task}}(1 + L(x)^{q/k}), \qquad \phi_H(x,a) \leq C_\phi(1 + L(x)).$$

The policy Π satisfies $\sup_x \mathrm{KL}(\Pi(\cdot \mid x)\|\nu) \leq C_{\mathrm{KL}} < \infty$. The features are linearly independent under μ_π for every admissible policy π.

The shared certificate channel and Assumption 5 give

$$0 \leq \epsilon_\pi(x;\theta,u) \leq 2\sup_a |w_u(x,a) - g_\theta(x,a)| \leq \bar{\epsilon}_\pi(1 + L(x)^{q/k}),$$

where one may take $\bar{\epsilon}_\pi = 2C_{\mathrm{task}}$.

Assumption 6 (Two-timescale step sizes and projections)**.** *The step sizes are positive and eventually nonincreasing, and satisfy*

$$\sum_t \alpha_t = \sum_t \beta_t = \infty, \qquad \sum_t (\alpha_t^2 + \beta_t^2) < \infty, \qquad \frac{\alpha_t}{\beta_t} \longrightarrow 0.$$

The projection sets $\Theta = \Theta_{\mathrm{task}} \times \Theta_H$ and U are compact and convex, with $\Theta_H \subset \mathbb{R}$ an interval; Γ_Θ is the Euclidean projection onto this product set. For SVPO with any fixed u, the mean critic dynamics admit an equilibrium in the interior of the projection set.

Define the state weight $\mathcal{V}(x) = 1 + L^m(x)$ and $\|f\|_{\mathcal{V}} = \sup_{x,a} |f(x,a)|/\mathcal{V}(x)$ for state–action functions. For state functions, the weighted supremum is over x; for vector-valued functions, use the Euclidean norm.

Assumption 7 (Policy and actor regularity)**.** *The function $w_u(x,a)$ is twice continuously differentiable in u. Its first two derivatives are bounded by $C(1 + L^{q/k})$ uniformly on U. There exists $C_{\pi,m} < \infty$ such that, for each f with $\|f\|_{\mathcal{V}} < \infty$, the policy averages satisfy*

$$\left\| \int f(\cdot,a)(\pi_\theta - \pi_{\theta'})(da \mid \cdot) \right\|_{\mathcal{V}} \leq C_{\pi,m}\|\theta - \theta'\|\|f\|_{\mathcal{V}},$$

$$\left\| \int f(\cdot, a)(\pi - \pi')(da \mid \cdot) \right\|_{\mathcal{V}} \leq C_{\pi,m}(\|u - u'\| + |\theta_H - \theta'_H|)\|f\|_{\mathcal{V}}.$$

Here π and π' are the practical policies at (u, θ_H) and (u', θ'_H). For Theorem 3, we additionally require policy-induced TD perturbations to be uniformly smaller than the frozen-policy TD decrease, in the precise sense of Appendix C.11.

Together with equation 31, these inequalities imply that the induced state and state–action kernels are Lipschitz in the parameters.

Assumption 8 (Quantitative minorization)**.** *For each weight $\mathcal{V}(x) = 1 + L^m(x)$ used in the convergence proof, there exist $\epsilon_m > 0$, $R_m < \infty$, and a probability measure ν_m on $\mathcal{X} \times \mathcal{A}$ such that every admissible exact or practical policy satisfies*

$$\mathsf{P}_\pi((x, a), \cdot) \geq \epsilon_m \nu_m(\cdot) \qquad \text{whenever } \mathcal{V}(x) \leq R_m.$$

The set $\{x : \mathcal{V}(x) \leq R_m\}$ contains the common drift set.

Note that geometric mixing already follows from Theorem 1. This assumption provides constants for error analysis along the policy sequence.

A.3 Certificate-relative equivalence for CPPI

To compare the expected next-step Lyapunov value under action a with the lowest value at the same state, we define

$$\ell_L(x) := \operatorname*{ess\,inf}_{a' \sim \nu} \mathsf{P}L(x, a'),$$
$$\chi_L(x, a) := \mathsf{P}L(x, a) - \ell_L(x).$$

The essential infimum is taken with respect to ν; since the policies considered below satisfy $\pi(\cdot \mid x) \ll \nu$, modifications on ν-null action sets do not affect the corresponding policy averages. For the converse characterization below, assume that there exist constants $0 < \underline{c} \leq \overline{c} < \infty$ and $C_R < \infty$ such that

$$\underline{c}\, s(x, a) - C_R(1 + L(x)^{q/k}) \leq \chi_L(x, a) \leq \overline{c}\, s(x, a) + C_R(1 + L(x)^{q/k}) \tag{16}$$

for every x and ν-almost every a.

Proposition 1 (Certificate-relative CPPI equivalence)**.** *Let $\pi(\cdot \mid x) \ll \nu$ for every x. Under equation 16, CPPI(m) holds if and only if, for every $1 \leq j \leq m$, there exists $D_j < \infty$ such that*

$$\int \chi_L(x, a)^j \pi(da \mid x) \leq D_j(1 + L(x)^{jq/k}) \qquad \text{for every } x. \tag{17}$$

B Proofs of the CPPI Results

B.1 Proof of Proposition 1

Proof. For $1 \leq j \leq m$, equation 16 and convexity of the jth power give

$$\begin{aligned} \chi_L^j &\leq (\overline{c}\, s + C_R(1 + L^{q/k}))^j \\ &\leq 2^{j-1}\overline{c}^{\,j} s^j + 2^{j-1} C_R^j (1 + L^{q/k})^j, \\ s^j &\leq \underline{c}^{-j}(\chi_L + C_R(1 + L^{q/k}))^j \\ &\leq 2^{j-1}\underline{c}^{-j} \left\{ \chi_L^j + C_R^j (1 + L^{q/k})^j \right\}, \end{aligned} \tag{18}$$

and

$$(1 + L^{q/k})^j \leq 2^{j-1}(1 + L^{jq/k}). \tag{19}$$

If π satisfies CPPI(m), integrating equation 18 and using equation 19 gives

$$\int \chi_L^j \, d\pi \leq \left(2^{j-1}\overline{c}^{\,j} C_j + 2^{2j-2} C_R^j \right) (1 + L^{jq/k}).$$

Conversely, equation 17 gives

$$\int s^j \, d\pi \leq 2^{j-1}\underline{c}^{-j} \left(D_j + 2^{j-1} C_R^j \right) (1 + L^{jq/k}).$$

These two implications hold for every $1 \leq j \leq m$ and prove the equivalence. □

B.2 Proof of Theorem 1

Proof. Fix $1 \le m \le \widetilde{m}$. For $x \notin \mathcal{C}$, equation 3 gives

$$0 \le \mathsf{P}L(x,a) \le (1-c_0)L(x) + s(x,a). \tag{20}$$

Hence, for any $\varepsilon > 0$ and $L(x) > R_\varepsilon$, Assumption 2, equation 20, and CPPI($\widetilde{m}$) imply

$$\begin{aligned}\mathsf{P}_\pi L^m(x) &\le (1+\varepsilon)\sum_{j=0}^{m}\binom{m}{j}(1-c_0)^{m-j}L(x)^{m-j}\int s(x,a)^j\pi(da \mid x)\\ &\le (1+\varepsilon)(1-c_0)^m L^m(x) + (1+\varepsilon)\sum_{j=1}^{m}\binom{m}{j}(1-c_0)^{m-j}C_j L(x)^{m-j}(1+L(x)^{jq/k}).\end{aligned} \tag{21}$$

Since $q/k = 1 - \Delta/k$ and $\Delta > 0$, for every $1 \le j \le m$, we have

$$L^{m-j}(1+L^{jq/k}) = L^{m-j} + L^{m-j\Delta/k} = o(L^m).$$

Thus every policy-dependent term in the sum in equation 21 is of strictly lower order than L^m.

Choose $\widetilde{\lambda}_{m,\pi} \in ((1-c_0)^m, 1)$, and then choose $\varepsilon > 0$ sufficiently small that

$$(1+\varepsilon)(1-c_0)^m < \widetilde{\lambda}_{m,\pi}.$$

Since the remaining terms in equation 21 are $o(L^m)$, there exists $R < \infty$, with $R \ge R_\varepsilon$ and $\mathcal{C} \subseteq \{x : L(x) \le R\}$, such that

$$\mathsf{P}_\pi L^m(x) \le \widetilde{\lambda}_{m,\pi} L^m(x) \qquad \text{whenever } L(x) > R.$$

Define $\widetilde{\mathcal{C}}_\pi := \{x : L(x) \le R\}$. This set is petite by Assumption 2. Define

$$\widetilde{b}_{m,\pi} := \sup_{x \in \widetilde{\mathcal{C}}_\pi} \left(\mathsf{P}_\pi L^m(x) - \widetilde{\lambda}_{m,\pi} L^m(x)\right)_+ .$$

The local moment bound in Assumption 2, together with equation 3 and CPPI($\widetilde{m}$), gives $\widetilde{b}_{m,\pi} < \infty$. Therefore

$$\mathsf{P}_\pi L^m(x) \le \widetilde{\lambda}_{m,\pi} L^m(x) + \widetilde{b}_{m,\pi}\mathbb{I}_{\widetilde{\mathcal{C}}_\pi}(x),$$

which proves equation 9.

Since $L^m \ge 1$, equation 9 together with Assumption 4 and the petiteness of $\widetilde{\mathcal{C}}_\pi$ implies positive Harris recurrence, a unique invariant measure μ_π, and L^m-geometric ergodicity by Meyn & Tweedie (2012, Theorem 15.0.1). Since

$$L^m(x) \le \mathcal{V}(x) := 1 + L^m(x) \le 2L^m(x),$$

the two weights are equivalent. Hence there exist $K_{m,\pi} < \infty$ and $\rho_{m,\pi} \in (0,1)$ such that, for every measurable $f : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$ with $\|f\|_{\mathcal{V}} < \infty$,

$$\begin{aligned}|\mathsf{P}_\pi^n f(x,a) - \mu_\pi(f)| &\le \int \mathsf{P}(dx' \mid x,a) K_{m,\pi}\rho_{m,\pi}^{n-1}\|f_\pi\|_{\mathcal{V}}\mathcal{V}(x')\\ &\le K_{m,\pi}C\rho_{m,\pi}^{n-1}\|f\|_{\mathcal{V}}\mathcal{V}(x),\end{aligned}$$

which proves equation 11.

Iterating equation 9 and using $\mathbb{I}_{\widetilde{\mathcal{C}}_\pi} \le 1$ gives

$$\begin{aligned}\mathsf{E}_x[L^m(X_t)] &\le \widetilde{\lambda}_{m,\pi}^t L^m(x) + \widetilde{b}_{m,\pi}\sum_{r=0}^{t-1}\widetilde{\lambda}_{m,\pi}^r = \widetilde{\lambda}_{m,\pi}^t L^m(x) + \frac{(1-\widetilde{\lambda}_{m,\pi}^t)\widetilde{b}_{m,\pi}}{1-\widetilde{\lambda}_{m,\pi}}\\ &\le \max\left\{L^m(x), \frac{\widetilde{b}_{m,\pi}}{1-\widetilde{\lambda}_{m,\pi}}\right\}.\end{aligned} \tag{22}$$

Averaging over any initial distribution with $\mathsf{E}[L^m(X_0)] < \infty$ yields

$$\sup_{t\geq 0} \mathsf{E}[L^m(X_t)] \leq \max\left\{\mathsf{E}[L^m(X_0)], \frac{\widetilde{b}_{m,\pi}}{1-\widetilde{\lambda}_{m,\pi}}\right\}.$$

The invariant-moment consequence of equation 9 gives $\mu_\pi(L^m) < \infty$. Using invariance of μ_π,

$$(1-\widetilde{\lambda}_{m,\pi})\mu_\pi(L^m) \leq \widetilde{b}_{m,\pi}\mu_\pi(\widetilde{\mathcal{C}}_\pi) \leq \widetilde{b}_{m,\pi},$$

and hence

$$\mu_\pi(L^m) \leq \frac{\widetilde{b}_{m,\pi}}{1-\widetilde{\lambda}_{m,\pi}}.$$

Together with equation 22, this proves equation 10.

For $m = 1$, $L \geq 1$ and $q/k < 1$ give $|c_\pi(x)| \leq C(1 + L(x)^{q/k}) \leq 2CL(x)$, so we have $\mu_\pi(|c_\pi|) < \infty$ and $\|c_\pi\|_{\mathcal{V}} < \infty$. Applying equation 11 with $f = c_\pi$ gives

$$\lim_{t\to\infty} \mathsf{E}_x[c_\pi(X_t)] = \mu_\pi(c_\pi).$$

Taking time averages therefore yields

$$J^\pi(x) = \lim_{T\to\infty} \frac{1}{T}\sum_{t=0}^{T-1} \mathsf{E}_x[c_\pi(X_t)] = \mu_\pi(c_\pi).$$

□

B.3 Proof of Corollary 1

Proof. By the definition of Λ_π and $q/k = 1 - \Delta/k$, whenever $L(x) \geq \Lambda_\pi$,

$$A_\pi L(x)^{q/k} = A_\pi L(x) L(x)^{-\Delta/k} \leq \frac{c_0}{4} L(x), \quad B_\pi \leq \frac{c_0}{4} L(x).$$

Therefore, for $x \notin \mathcal{C}$ with $L(x) \geq \Lambda_\pi$, the grade-one drift bound gives

$$\begin{aligned}\mathsf{P}_\pi L(x) &\leq (1-c_0)L(x) + A_\pi L(x)^{q/k} + B_\pi \\ &\leq \left(1 - \frac{c_0}{2}\right) L(x).\end{aligned}$$

For any $1 \leq m \leq \widetilde{m}$ and $\gamma > 1$, Markov's inequality gives

$$\mu_\pi(L > \gamma\Lambda_\pi) = \mu_\pi(L^m > \gamma^m \Lambda_\pi^m) \leq \frac{\mu_\pi(L^m)}{\gamma^m \Lambda_\pi^m}.$$

Applying equation 10 then yields

$$\mu_\pi(L > \gamma\Lambda_\pi) \leq \frac{\widetilde{b}_{m,\pi}}{(1-\widetilde{\lambda}_{m,\pi})\gamma^m \Lambda_\pi^m},$$

which proves the claim. □

C SVPO Derivations and Convergence Proofs

C.1 Policy-space properties of Gibbs improvement and Tilt

Proposition 2 (Gibbs and Tilt policy-space properties)**.** *At every state where the relevant normalizers and expectations are finite:*

1. Gibbs optimality. *The Gibbs policy in equation 5 is the unique global minimizer of the statewise free energy in equation 6.*

2. Certificate preservation and minimum KL. *Let ϱ satisfy* $\inf_x \varrho\{a : s(x,a) \le \xi_0 \mid x\} \ge p_0$ *and let* $\eta > 0$. *Then*

$$\mathrm{Tilt}_\eta(\varrho)(da \mid x) = \frac{e^{-\eta\phi_H(x,a)/\tau}\,\varrho(da \mid x)}{\int e^{-\eta\phi_H(x,a')/\tau}\,\varrho(da' \mid x)} \tag{23}$$

satisfies CPPI(∞) with rate η/τ. With $b_\eta(x) := \mathbb{E}_{\mathrm{Tilt}_\eta(\varrho)}[s(x,A)]$, this policy is also the unique minimizer of $\mathrm{KL}(\pi(\cdot \mid x)\|\varrho(\cdot \mid x))$ *subject to* $\mathbb{E}_\pi[s(x,A)] \le b_\eta(x)$.

3. Load monotonicity. *The load satisfies*

$$\frac{\partial b_\eta(x)}{\partial \eta} = -\frac{1}{\tau}\,\mathrm{Var}_{\mathrm{Tilt}_\eta(\varrho)}[s(x,A)] \le 0.$$

Hence $b_\eta(x)$ is strictly decreasing wherever the variance is positive.

C.2 Proof of Proposition 2

Proof. Substituting the Gibbs density in equation 5 gives

$$\mathcal{G}_x(\pi;\theta) - \mathcal{G}_x(\pi_\theta;\theta) = \tau\mathrm{KL}(\pi(\cdot \mid x)\|\pi_\theta(\cdot \mid x)).$$

Since the KL divergence is nonnegative and vanishes only when $\pi(\cdot \mid x) = \pi_\theta(\cdot \mid x)$, the Gibbs policy is the unique minimizer.

For Tilt, let $\pi_\eta = \mathrm{Tilt}_\eta(\varrho)$. Since $\phi_H(x,a) = s(x,a) + u^*(x)$ and $u^*(x)$ is independent of a, the state-only term cancels under normalization. Hence

$$\begin{aligned}\pi_\eta(s \ge s_0 \mid x) &= \frac{\int_{\{s \ge s_0\}} e^{-\eta s/\tau}\,d\varrho}{\int e^{-\eta s/\tau}\,d\varrho} \le \frac{e^{-\eta s_0/\tau}}{\varrho(s \le \xi_0 \mid x)e^{-\eta\xi_0/\tau}} \\ &\le p_0^{-1}e^{\eta\xi_0/\tau}e^{-\eta s_0/\tau}.\end{aligned}$$

Since $1 + L(x)^{q/k} \ge 1$, this satisfies CPPI(∞).

For the minimum-KL property, define $Z_\eta(x) := \int e^{-\eta s/\tau}\,d\varrho$. Then

$$\log\frac{d\pi_\eta}{d\varrho} = -\frac{\eta}{\tau}s - \log Z_\eta(x).$$

For any $\pi \ll \varrho$ satisfying $\mathbb{E}_\pi[s] \le b_\eta(x)$,

$$\mathrm{KL}(\pi\|\varrho) - \mathrm{KL}(\pi_\eta\|\varrho) = \mathrm{KL}(\pi\|\pi_\eta) + \frac{\eta}{\tau}\left(b_\eta(x) - \mathbb{E}_\pi[s]\right) \ge 0.$$

Equality holds only when $\pi = \pi_\eta$, which proves uniqueness.

Finally,

$$b_\eta(x) = -\tau\,\partial_\eta \log Z_\eta(x), \quad \frac{\partial b_\eta(x)}{\partial \eta} = -\frac{1}{\tau}\,\mathrm{Var}_{\pi_\eta}[s(x,A)] \le 0.$$

□

On a compact coefficient interval where the variance is positive, each target in the interior of the attainable load range has a unique interior multiplier. Clipping selects the corresponding endpoint for targets outside this range.

C.3 Complete online recursion

We initialize $\theta_0 \in \Theta$ and $u_0 \in U$, with $\mathsf{E}[|\rho_0|^{2m} + L^{2m}(X_0)] < \infty$ for grade m. For the kernel critic defined in the main text, the semi-gradient is

$$\nabla_\theta \widehat{Q}_\theta(x,a) = \begin{pmatrix} \psi_K(x,a) \\ \eta'(\theta_H)\phi_H(x,a) \end{pmatrix}.$$

For each transition, draw $\widetilde{A}_{t+1} \sim \pi_{u_t}(\cdot \mid X_{t+1})$ from the same pre-update raw policy and set

$$\Delta_t = c(X_t, A_t) - \rho_t + \widehat{Q}_{\theta_t}(X_{t+1}, \widetilde{A}_{t+1}) - \widehat{Q}_{\theta_t}(X_t, A_t).$$

The average-cost and critic updates are

$$\rho_{t+1} = \rho_t + \beta_t(c(X_t, A_t) - \rho_t),$$
$$\theta_{t+1} = \Gamma_\Theta\Big[\theta_t + \beta_t \Delta_t \nabla_\theta \widehat{Q}_{\theta_t}(X_t, A_t)\Big].$$

Using equation 8 with the current parameters, the projected actor update is

$$u_{t+1} = \Gamma_U\Big(u_t - \alpha_t \widehat{\nabla}_u \mathcal{L}_{\text{actor}}(u_t; \theta_t)\Big).$$

The exact realization samples from equation 5 and uses the same fast updates without a separate actor recursion.

C.4 Proof of Theorem 2

Proof. For the practical policy, cancellation of the state-only term $u^*(x)$ gives

$$\pi_u(da \mid x) = \frac{\exp[-(w_u(x,a) + \eta(\theta_H)s(x,a))/\tau]\nu(da)}{\int \exp[-(w_u(x,a') + \eta(\theta_H)s(x,a'))/\tau]\nu(da')}.$$

Assumptions 3 and 5 imply

$$\int_{\{s \geq s_0\}} e^{-(w_u + \eta(\theta_H)s)/\tau}\, d\nu \leq e^{C_{\text{task}}(1+L(x)^{q/k})/\tau} e^{-\eta(\theta_H)s_0/\tau},$$
$$\int e^{-(w_u + \eta(\theta_H)s)/\tau}\, d\nu \geq \int_{\{s \leq \xi_0\}} e^{-(w_u + \eta(\theta_H)s)/\tau}\, d\nu \geq p_0 e^{-C_{\text{task}}(1+L(x)^{q/k})/\tau} e^{-\eta(\theta_H)\xi_0/\tau}.$$

Therefore,

$$\pi_u(s \geq s_0 \mid x) \leq p_0^{-1} \exp\left[\frac{2C_{\text{task}}(1 + L(x)^{q/k}) + \eta(\theta_H)\xi_0}{\tau} - \frac{\eta(\theta_H)}{\tau} s_0\right]. \tag{24}$$

Define

$$\gamma := \frac{\eta(\theta_H)}{\tau}, \quad \theta_s := \log p_0^{-1} + \frac{2C_{\text{task}} + \eta(\theta_H)\xi_0}{\tau}.$$

Since $1 + L(x)^{q/k} \geq 1$, equation 24 implies

$$\pi_u(s \geq s_0 \mid x) \leq \exp\Big[\theta_s(1 + L(x)^{q/k}) - \gamma s_0\Big].$$

Hence π_u satisfies CPPI(∞) with $\gamma = \frac{\eta(\theta_H)}{\tau} \geq \frac{\epsilon_0}{\tau}$. The exact Gibbs policy satisfies the same bound because g_θ and w_u obey the same growth bound in Assumption 5.

Moreover, CPPI(∞) yields CPPI(m) for every finite m. For every $j \geq 1$, the tail-integral identity and the exponential bound above give

$$\mathbb{E}[s^j \mid x] = j \int_0^\infty s_0^{j-1} \Pr(s \geq s_0 \mid x)\, ds_0 \leq C_j \left[1 + \left(\frac{\theta_s}{\gamma}(1 + L(x)^{q/k})\right)^j\right]$$
$$\leq C'_j(1 + L(x)^{jq/k}),$$

where C_j and C'_j are finite constants. Thus CPPI(∞), and therefore every finite CPPI grade, holds for both policies. □

C.5 Drift coefficients and sufficient certificate strength

Under Assumptions 1, 5, and 7, the free-energy comparison gives

$$\mathcal{G}_x(\pi_u;\theta) \le \mathcal{G}_x(\Pi;\theta) + \bar{\epsilon}_\pi(1+L^{q/k}).$$

Expanding both sides and dropping the nonnegative term $\tau\mathrm{KL}(\pi_u(\cdot \mid x)\|\nu)$ yields

$$\eta(\theta_H)\left(\int \phi_H\, d\pi_u - \int \phi_H\, d\Pi\right) \le \int g_\theta\, d\Pi - \int g_\theta\, d\pi_u + \tau\mathrm{KL}(\Pi(\cdot \mid x)\|\nu) + \bar{\epsilon}_\pi(1+L^{q/k}).$$

Define

$$A_{\pi_u} := \frac{2C_{\mathrm{task}} + \bar{\epsilon}_\pi}{\eta(\theta_H)}, \qquad B_{\pi_u} := \frac{2C_{\mathrm{task}} + \tau C_{\mathrm{KL}} + \bar{\epsilon}_\pi}{\eta(\theta_H)}. \tag{25}$$

Then

$$\begin{aligned}\int \phi_H\, d\pi_u - \mathcal{D} &\le \int \phi_H\, d\Pi - \mathcal{D} + \frac{2C_{\mathrm{task}}(1+L^{q/k}) + \tau C_{\mathrm{KL}} + \bar{\epsilon}_\pi(1+L^{q/k})}{\eta(\theta_H)} \\ &\le -c_0 L + b_0 \mathbb{I}_{\mathcal{C}} + A_{\pi_u} L^{q/k} + B_{\pi_u}.\end{aligned}$$

Thus

$$\int \phi_H(x,a)\pi_u(da \mid x) - \mathcal{D}(x) \le -c_0 L(x) + A_{\pi_u} L(x)^{q/k} + B_{\pi_u} + b_0 \mathbb{I}_{\mathcal{C}}(x).$$

For the exact Gibbs realization, set $\bar{\epsilon}_\pi = 0$.

Now fix $\Lambda_* > 1$. By equation 12, $\Lambda_{\pi_u} \le \Lambda_*$ is ensured by $A_{\pi_u} \le \frac{c_0}{4}\Lambda_*^{\Delta/k}$ and $B_{\pi_u} \le \frac{c_0}{4}\Lambda_*$. Substituting equation 25 into equation 12 gives

$$\eta(\theta_H) \ge \eta_{\min}(\Lambda_*) := \max\left\{\frac{4(2C_{\mathrm{task}} + \bar{\epsilon}_\pi)}{c_0 \Lambda_*^{\Delta/k}}, \frac{4(2C_{\mathrm{task}} + \tau C_{\mathrm{KL}} + \bar{\epsilon}_\pi)}{c_0 \Lambda_*}\right\}. \tag{26}$$

C.6 Necessity and sufficiency of a positive certificate floor

Consider the practical policy family

$$\pi_u(da \mid x) \propto \exp\left[-\frac{w_u(x,a) + \eta(\theta_H)\phi_H(x,a)}{\tau}\right]\nu(da),$$

over the admissible parameters $(u,\theta) \in \mathcal{U} \times \Theta$, and suppose

$$0 \le \eta(\theta_H) \le \bar{\eta} < \infty.$$

By Assumption 5,

$$\sup_a w_u(x,a) - \inf_a w_u(x,a) \le 2C_{\mathrm{task}}(1 + L(x)^{q/k}) \tag{27}$$

uniformly over the admissible parameters.

For the necessity direction, assume additionally that there exist constants $c_-, c_+, p_1 > 0$, states x_n with $L(x_n) \to \infty$, and measurable sets R_n such that

$$\nu(R_n) \ge p_1, \qquad c_- L(x_n) \le s(x_n,a) \le c_+ L(x_n), \qquad a \in R_n. \tag{28}$$

Corollary 4 (Necessity and sufficiency of a positive certificate floor)**.** *Under Assumption 3 and equation 28, the practical policy family admits common CPPI(∞) constants if and only if its certificate coefficients have a common positive lower bound.*

C.7 Proof of Corollary 4

Proof. Since $\phi_H(x,a) = s(x,a) + u^*(x)$ and $u^*(x)$ is independent of the action, the factor $\exp[-\eta(\theta_H)u^*(x)/\tau]$ cancels under normalization. Hence every policy in the family can equivalently be written as

$$\pi_u(da \mid x) \propto \exp\left[-\frac{w_u(x,a) + \eta(\theta_H)s(x,a)}{\tau}\right]\nu(da).$$

Suppose first that the policy family admits common CPPI(∞) constants (θ_s, γ). Since $s \geq 0$ and $\eta(\theta_H) \geq 0$,

$$\int e^{-(w_u(x_n,a)+\eta(\theta_H)s(x_n,a))/\tau}\nu(da) \leq e^{-\inf_a w_u(x_n,a)/\tau}.$$

On R_n, equation 28 gives $s(x_n,a) \leq c_+L(x_n)$. Therefore, using equation 27 and $\nu(R_n) \geq p_1$,

$$\pi_u(R_n \mid x_n) \geq p_1 \exp\left[-\frac{2C_{\text{task}}(1+L(x_n)^{q/k}) + \eta(\theta_H)c_+L(x_n)}{\tau}\right].$$

Moreover, equation 28 implies $R_n \subseteq \{a : s(x_n,a) \geq c_-L(x_n)\}$. Hence the common CPPI($\infty$) bound gives

$$\begin{aligned}\pi_u(R_n \mid x_n) &\leq \pi_u(s \geq c_-L(x_n) \mid x_n)\\ &\leq \exp\left[\theta_s(1+L(x_n)^{q/k}) - \gamma c_-L(x_n)\right].\end{aligned}$$

Combining the lower and upper bounds and taking logarithms yields

$$\eta(\theta_H) \geq \frac{\tau\gamma c_-}{c_+} + \frac{\tau\log p_1}{c_+L(x_n)} - \frac{(2C_{\text{task}} + \tau\theta_s)(1+L(x_n)^{q/k})}{c_+L(x_n)}.$$

Since $q/k < 1$ and $L(x_n) \to \infty$, letting $n \to \infty$ gives

$$\eta(\theta_H) \geq \frac{\tau\gamma c_-}{c_+} > 0.$$

The right-hand side is independent of the selected policy. Hence the family has a common positive certificate floor.

Conversely, suppose there exists $\eta_0 > 0$ such that $\eta(\theta_H) \geq \eta_0$ for every admissible policy in the family. By assumption, $\eta(\theta_H) \leq \bar{\eta}$ as well. For any $s_0 \geq 0$, Assumption 3 gives a set $\{a : s(x,a) \leq \xi_0\}$ with ν-mass at least p_0. Hence

$$\begin{aligned}\pi_u(s \geq s_0 \mid x) &= \frac{\int_{\{s\geq s_0\}} e^{-(w_u+\eta(\theta_H)s)/\tau}\,d\nu}{\int e^{-(w_u+\eta(\theta_H)s)/\tau}\,d\nu}\\ &\leq \frac{e^{-\inf_a w_u(x,a)/\tau}e^{-\eta(\theta_H)s_0/\tau}}{p_0 e^{-\sup_a w_u(x,a)/\tau}e^{-\eta(\theta_H)\xi_0/\tau}}\\ &\leq p_0^{-1}\exp\left[\frac{2C_{\text{task}}(1+L(x)^{q/k}) + \bar{\eta}\xi_0}{\tau} - \frac{\eta_0}{\tau}s_0\right],\end{aligned}$$

where the last inequality uses equation 27, $\eta(\theta_H) \leq \bar{\eta}$, and $\eta(\theta_H) \geq \eta_0$.

Thus the last bound is of the CPPI(∞) form uniformly over the policy family, with common rate $\gamma = \eta_0/\tau$ and a finite θ_s independent of the policy. This proves the converse. □

C.8 Proof of Corollary 2

Proof. We first show that the CPPI constants can be chosen uniformly over the exact Gibbs and practical SVPO policy families. By equation 24,

$$\pi(s \geq s_0 \mid x) \leq p_0^{-1}\exp\left[\frac{2C_{\text{task}}(1+L(x)^{q/k})}{\tau} - \frac{\eta(\theta_H)}{\tau}(s_0 - \xi_0)\right].$$

For $s_0 \geq \xi_0$, the positive floor $\eta(\theta_H) \geq \epsilon_0$ gives

$$\pi(s \geq s_0 \mid x) \leq p_0^{-1} \exp\left[\frac{2C_{\text{task}}(1 + L(x)^{q/k})}{\tau} - \frac{\epsilon_0}{\tau}(s_0 - \xi_0)\right].$$

For $0 \leq s_0 < \xi_0$, the same CPPI(∞) form follows from $\pi(s \geq s_0 \mid x) \leq 1$. Hence both policy families satisfy CPPI(∞) with common constants, including the common rate $\gamma = \epsilon_0/\tau$.

The tail-integration argument in the proof of Theorem 2 therefore gives, for $1 \leq j \leq \widetilde{m}$,

$$\int s(x,a)^j \pi(da \mid x) \leq C_j(1 + L(x)^{jq/k}),$$

where the constants C_j are uniform over both policy families. Applying the construction in Theorem 1 with these common constants gives, for every $1 \leq m \leq \widetilde{m}$, $\lambda_m \in (0,1)$, $b_m < \infty$, and an L-bounded set $\mathcal{C}_m$, common to both policy families, such that

$$\mathsf{P}_\pi L^m \leq \lambda_m L^m + b_m \mathbb{I}_{\mathcal{C}_m}. \tag{29}$$

At grade one, choosing the common coefficients $A_\pi = B_\pi = C_1$ in equation 12 also gives a common tail threshold, so the conclusion of Corollary 1 holds uniformly.

Marginalizing the action coordinate in Assumption 8 gives a common local minorization for the state chain. Together with equation 29, this yields constants $K_m < \infty$ and $\rho_m \in (0,1)$, independent of the policy, such that

$$|\mathsf{P}_\pi^n f(x) - \mu_\pi(f)| \leq K_m \rho_m^n \|f\|_{\mathcal{V}} \mathcal{V}(x) \tag{30}$$

for every state function f with $\|f\|_{\mathcal{V}} < \infty$.

For the corresponding state–action bound, Assumptions 1 and 5, together with $\mathcal{D} \geq 0$, give

$$\begin{aligned}\mathsf{P}L(x,a) &\leq L(x) + \phi_H(x,a) + b_1 \\ &\leq C(1 + L(x)).\end{aligned}$$

Assumption 2 then implies

$$(\mathsf{P}\mathcal{V})(x,a) = 1 + \mathsf{P}L^m(x,a) \leq C\mathcal{V}(x) \tag{31}$$

uniformly in a. For a state–action function f with $\|f\|_{\mathcal{V}} < \infty$, let

$$f_\pi(x) = \int f(x,a)\pi(da \mid x).$$

Then $\|f_\pi\|_{\mathcal{V}} \leq \|f\|_{\mathcal{V}}$, and for $n \geq 1$,

$$\begin{aligned}|\mathsf{P}_\pi^n f(x,a) - \mu_\pi(f)| &\leq \int \mathsf{P}(dx' \mid x,a) K_m \rho_m^{n-1} \|f_\pi\|_{\mathcal{V}} \mathcal{V}(x') \\ &\leq K_m C \rho_m^{n-1} \|f\|_{\mathcal{V}} \mathcal{V}(x),\end{aligned}$$

where equation 30 and equation 31 give the two inequalities. After enlarging K_m if necessary, this has the same geometric form with ρ_m^n.

Let $\mathcal{G}_t$ contain the history through X_t and the current parameters, before sampling A_t, and set $\mathcal{F}_t = \mathcal{G}_t \vee \sigma(A_t)$, before sampling $(X_{t+1}, \widetilde{A}_{t+1})$. Finally, equation 29 holds for every policy generated along the adaptive SVPO path. Therefore,

$$\begin{aligned}\mathsf{E}[L^m(X_{t+1}) \mid \mathcal{G}_t] &= \mathsf{P}_{\pi_t} L^m(X_t) \\ &\leq \lambda_m L^m(X_t) + b_m.\end{aligned}$$

Taking expectations and iterating gives

$$\mathsf{E}[L^m(X_t)] \leq \lambda_m^t \mathsf{E}[L^m(X_0)] + \frac{b_m(1 - \lambda_m^t)}{1 - \lambda_m},$$

and hence

$$\sup_{t \geq 0} \mathsf{E}[L^m(X_t)] \leq \max\left\{\mathsf{E}[L^m(X_0)], \frac{b_m}{1 - \lambda_m}\right\} < \infty. \tag{32}$$

□

C.9 Proof of Lemma 1

Proof. Use the grade m and weight $\mathcal{V}(x) = 1 + L^m(x)$ defined in the main text. Compactness of Θ and Assumption 5 give

$$\sup_{\theta\in\Theta}(|\eta(\theta_H)|+|\eta'(\theta_H)|+|\eta''(\theta_H)|) < \infty,$$
$$|\widehat{Q}_\theta(z)|+\|\nabla_\theta\widehat{Q}_\theta(z)\|+\|\nabla^2_\theta\widehat{Q}_\theta(z)\| \le C(1+L(x)).$$

Consequently, for every $\rho \in \mathbb{R}$,

$$|\Delta(\rho,\theta;z,z')| \le C(2+L(x)+L(x')+|\rho|),$$
$$\|H_c(z,z';\rho,\theta)\| \le C\left(\mathcal{V}(x)+\mathcal{V}(x')+|\rho|(1+L(x))\right).$$

Rewriting the practical policy with $s = \phi_H - u^*$ gives

$$\nabla_u \log\frac{d\pi_u}{d\nu}(a \mid x) = -\frac{1}{\tau}\left(\nabla_u w_u(x,a) - \int \nabla_u w_u(x,a')\pi_u(da' \mid x)\right).$$

The shared certificate term cancels in $\widehat{Q}_\theta + \tau\log(d\pi_u/d\nu)$. The normalizer bounds used in the proof of Theorem 2, together with $0 \le u^*(x) \le C(1+L(x))$, therefore give

$$\left\|\nabla_u \log\frac{d\pi_u}{d\nu}(a \mid x)\right\| \le C(1+L(x)^{q/k}),$$
$$\left|\widehat{Q}_\theta(x,a) + \tau\log\frac{d\pi_u}{d\nu}(a \mid x)\right| \le C(1+L(x)),$$
$$\|H_a(z;u,\theta)\| \le C(1+L(x))(1+L(x)^{q/k}) \le C\mathcal{V}(x),$$

where the last inequality uses $1 + q/k < 2 \le m$.

Applying equation 31 at grades m and $2m$ controls the update fields and their second moments:

$$\|h_c(\cdot;u,\rho,\theta)\|_{\mathcal{V}} \le C(1+|\rho|), \quad \|h_a(\cdot;u,\theta)\|_{\mathcal{V}} \le C,$$
$$\int \|H_c(z,z';\rho,\theta)\|^2 \mathsf{P}_{\pi_u}(z,dz') + \|H_a(z;u,\theta)\|^2 \le C(\mathcal{V}(x)^2 + |\rho|^2(1+L(x))^2).$$

The critic field is affine in ρ:

$$h_c(z;u,\rho,\theta) = h_c(z;u,0,\theta) - \rho\begin{pmatrix}1\\ \nabla_\theta\widehat{Q}_\theta(z)\end{pmatrix}.$$

The bounds above and the bounded second derivatives of $\widehat{Q}_\theta$ control the dependence of H_c on (ρ,θ), with the affine decomposition giving the factor $1 + |\rho|+|\rho'|$. Assumption 7 controls the corresponding change of the transition kernel with the policy parameters. Together with the first two derivative bounds on w_u, these estimates give equation 14.

It remains to verify the same regularity for the invariant means. For either conditional field, let f and f' denote its values at two parameter choices, with policies π_u and $\pi_{u'}$. Then

$$\mu_{\pi_u}(f) - \mu_{\pi_{u'}}(f') = \mu_{\pi_u}(f-f') + (\mu_{\pi_u} - \mu_{\pi_{u'}})(f'),$$
$$(\mu_{\pi_u} - \mu_{\pi_{u'}})(f') = \mu_{\pi_u}(\mathsf{P}_{\pi_u} - \mathsf{P}_{\pi_{u'}})\sum_{n=0}^{\infty}\left(\mathsf{P}^n_{\pi_{u'}}f' - \mu_{\pi_{u'}}(f')\right).$$

Assumption 7 controls the kernel difference, and Corollary 2 gives a common invariant $\mathcal{V}$-moment bound. Combining these estimates with the field bounds above gives the corresponding Lipschitz bounds for $\bar{h}_c$ and $\bar{h}_a$. □

C.10 Proof of Lemma 2

Proof. Fix admissible parameters and a field f with $\|f\|_{\mathcal{V}} < \infty$. The uniform geometric bound in Corollary 2 implies that

$$\widehat{f} = \sum_{n=0}^{\infty}\left(\mathsf{P}^n_{\pi_u}f - \mu_{\pi_u}(f)\right)$$

converges in $\mathcal{V}$-norm. We have $\mu_{\pi_u}(\widehat{f}) = 0$, and hence

$$\widehat{f} - \mathsf{P}_{\pi_u}\widehat{f} = f - \mu_{\pi_u}(f), \qquad \|\widehat{f}\|_{\mathcal{V}} \le C_{\mathcal{V}}\|f\|_{\mathcal{V}}.$$

For a second field f' and parameter choice (u', θ'), let $\widehat{f'}$ denote the corresponding centered Poisson solution. The invariant-measure perturbation identity in the preceding proof, uniform mixing and Assumption 7 give

$$\|\widehat{f} - \widehat{f'}\|_{\mathcal{V}} \le C\left(\|f - f'\|_{\mathcal{V}} + \|(u,\theta) - (u',\theta')\|\,\|f'\|_{\mathcal{V}}\right).$$

The same estimates hold for the state-chain kernel and for the weight $1 + L$, using the corresponding bounds in Corollary 2.

By Assumption 6, choose t_0 so that $0 < \beta_t \le 1$ for $t \ge t_0$. Convexity of $|\cdot|^{2m}$ gives

$$|\rho_{t+1}|^{2m} \le (1-\beta_t)|\rho_t|^{2m} + \beta_t |c(X_t, A_t)|^{2m}. \tag{33}$$

The initialization in Appendix C.3, Corollary 2 at grade $2m$ and Assumption 5 imply

$$\begin{aligned}\sup_t \mathsf{E}|c(X_t, A_t)|^{2m} &\le C \sup_t \mathsf{E}[1 + L(X_t)^{2mq/k}] \\ &\le C \sup_t \mathsf{E}[1 + L(X_t)^{2m}] < \infty.\end{aligned}$$

Taking expectations in equation 33 and iterating gives

$$\sup_{t\ge 0} \mathsf{E}|\rho_t|^{2m} \le \max\left\{\max_{t \le t_0} \mathsf{E}|\rho_t|^{2m}, \sup_t \mathsf{E}|c(X_t, A_t)|^{2m}\right\} < \infty. \tag{34}$$

Young's inequality with conjugate exponents m and $m/(m-1)$ gives

$$\begin{aligned}|\rho_t|^2(1 + L(X_t))^2 &\le \frac{|\rho_t|^{2m}}{m} + \frac{m-1}{m}(1 + L(X_t))^{2m/(m-1)} \\ &\le C(|\rho_t|^{2m} + \mathcal{V}(X_t)^2),\end{aligned}$$

where $m \ge 2$ bounds the state exponent by $2m$. Together with equation 34 and equation 32, this yields

$$\sup_{t\ge 0} \mathsf{E}\left[\mathcal{V}(X_t)^2 + |\rho_t|^2(1 + L(X_t))^2\right] < \infty. \tag{35}$$

For a state–action field f, define $f_\pi(x) := \int f(x,a)\pi(da \mid x)$. Assumption 7 transfers the field bounds in Lemma 1 to these statewise policy averages. Let $\widehat{h}_c$ and $\widehat{h}_a$ be the centered state-chain Poisson solutions associated with $(h_c)_{\pi_u}$ and $(h_a)_{\pi_u}$, respectively. The coefficient of ρ in h_c is $(1, \nabla_\theta \widehat{Q}_\theta)^\top$; after centering, its constant first coordinate vanishes. Let $\widehat{r}$ therefore denote the centered Poisson solution for the policy average of $(0, \nabla_\theta \widehat{Q}_\theta)^\top$. The affine decomposition from the preceding proof gives $\widehat{h}_c(\cdot; u, \rho, \theta) = \widehat{h}_c(\cdot; u, 0, \theta) - \rho\widehat{r}(\cdot; u, \theta)$. The resolvent and kernel bounds with weight $1 + L$ give $\|\widehat{r}(\cdot; u, \theta)\|_{1+L} \le C$ and $\|\widehat{r}(\cdot; u, \theta) - \widehat{r}(\cdot; u', \theta')\|_{1+L} \le C\|(u,\theta) - (u',\theta')\|$. Together with the $\mathcal{V}$-bounds for $\widehat{h}_c(\cdot; u, 0, \theta)$ and $\widehat{h}_a$, these imply

$$\begin{aligned}\|\widehat{h}_c(x; u, \rho, \theta)\| &\le C(\mathcal{V}(x) + |\rho|(1 + L(x))), \\ \|\widehat{h}_a(x; u, \theta)\| &\le C\mathcal{V}(x), \\ \|\widehat{h}_c(x; u, \rho, \theta) - \widehat{h}_c(x; u', \rho', \theta')\| &\le C\|(u,\theta) - (u',\theta')\|(\mathcal{V}(x) + |\rho'|(1 + L(x))) \\ &\quad + C|\rho - \rho'|(1 + L(x)), \\ \|\widehat{h}_a(x; u, \theta) - \widehat{h}_a(x; u', \theta')\| &\le C\|(u,\theta) - (u',\theta')\|\mathcal{V}(x).\end{aligned}$$

Define $\pi_t = \pi_{u_t}$ for the practical policy. Take (ω_t, f_t) to be either $(\beta_t, h_c(\cdot; u_t, \rho_t, \theta_t))$ or $(\alpha_t, h_a(\cdot; u_t, \theta_t))$, and let $\widehat{f}_t$ be the corresponding centered state-chain Poisson solution:

$$\widehat{f}_t - \mathsf{P}_{\pi_t}\widehat{f}_t = f_{t,\pi_t} - \mu_{\pi_t}(f_t).$$

Since

$$\mathsf{E}[f_t(Z_t) \mid \mathcal{G}_t] = f_{t,\pi_t}(X_t), \qquad \mathsf{E}[\widehat{f}_t(X_{t+1}) \mid \mathcal{G}_t] = \mathsf{P}_{\pi_t}\widehat{f}_t(X_t),$$

the centered field admits the exact decomposition

$$f_t(Z_t) - \mu_{\pi_t}(f_t) = f_t(Z_t) - f_{t,\pi_t}(X_t) + \widehat{f}_t(X_t) - \widehat{f}_t(X_{t+1}) + \widehat{f}_t(X_{t+1}) - \mathsf{E}[\widehat{f}_t(X_{t+1}) \mid \mathcal{G}_t]. \tag{36}$$

The first and last terms on the right are martingale differences, while the middle term is handled by summation by parts.

Lemma 1, the one-step bound in equation 31, equation 35, and Assumption 6 give

$$\sum_t \omega_t^2 \mathsf{E}\left[\|f_t(Z_t) - f_{t,\pi_t}(X_t)\|^2 + \|\widehat{f}_t(X_{t+1})\|^2\right] < \infty.$$

Hence both step-size-weighted martingale terms in equation 36 converge almost surely, and square summability also gives $\omega_n \widehat{f}_n(X_{n+1}) \to 0$ almost surely. Summation by parts gives

$$\begin{aligned}
&\sum_{t=0}^{n} \omega_t(\widehat{f}_t(X_t) - \widehat{f}_t(X_{t+1})) \\
&\quad = \omega_0\widehat{f}_0(X_0) - \omega_n\widehat{f}_n(X_{n+1}) + \sum_{t=1}^{n}(\omega_t - \omega_{t-1})\widehat{f}_{t-1}(X_t) + \sum_{t=1}^{n}\omega_t(\widehat{f}_t - \widehat{f}_{t-1})(X_t).
\end{aligned} \tag{37}$$

The Poisson-solution and one-step moment bounds, equation 35, and eventual monotonicity of the step sizes imply

$$\sum_{t\geq 1} |\omega_t - \omega_{t-1}| \mathsf{E}\|\widehat{f}_{t-1}(X_t)\| < \infty.$$

Projection nonexpansiveness and the update-field bounds give

$$\begin{aligned}
\|(u_t, \theta_t) - (u_{t-1}, \theta_{t-1})\| &\leq C(\beta_{t-1} + \alpha_{t-1})(\mathcal{V}(X_{t-1}) + \mathcal{V}(X_t) + |\rho_{t-1}|(1 + L(X_{t-1}))), \\
|\rho_t - \rho_{t-1}| &\leq C\beta_{t-1}(1 + L(X_{t-1}) + |\rho_{t-1}|).
\end{aligned}$$

For either field, the parameter-Lipschitz bounds and Young's inequality yield

$$\|(\widehat{f}_t - \widehat{f}_{t-1})(X_t)\| \leq C(\beta_{t-1} + \alpha_{t-1})(|\rho_{t-1}|^{2m} + \mathcal{V}(X_{t-1})^2 + \mathcal{V}(X_t)^2).$$

Taking expectations and using equation 34, equation 32, and the square-summable step sizes gives

$$\sum_{t\geq 1} \omega_t \mathsf{E}\|(\widehat{f}_t - \widehat{f}_{t-1})(X_t)\| \leq C \sum_{t\geq 1} \omega_t(\beta_{t-1} + \alpha_{t-1}) < \infty.$$

Equations 36–37 therefore prove both almost-sure convergence statements. Replacing π_u by π_θ gives the exact-policy statement used below. □

C.11 SARSA-control regularity for Theorem 3

Let $z = (x, a)$ and $z' = (x', a')$. Define:

$$\xi(z) := \begin{pmatrix} \psi_K(z) \\ \phi_H(z) \end{pmatrix}, \qquad y(\rho, \theta) := \begin{pmatrix} \rho \\ \theta_{\text{task}} \\ \eta(\theta_H) \end{pmatrix}.$$

For an admissible policy π, define $c(z) = c(x, a)$ and

$$\begin{aligned}
b_\pi &:= \begin{pmatrix} J^\pi \\ \int c(z)\xi(z)\mu_\pi(dz) \end{pmatrix}, \\
\mathbf{A}_\pi &:= \begin{pmatrix} 1 & 0 \\ \int \xi(z)\mu_\pi(dz) & \iint \xi(z)(\xi(z) - \xi(z'))^\top \mu_\pi(dz)\mathsf{P}_\pi(z, dz') \end{pmatrix}.
\end{aligned} \tag{38}$$

With I denoting the identity matrix, the mean critic update in equation 13 can be written as

$$\bar{h}_c = \mathrm{diag}(1, I, \eta'(\theta_H))(b_\pi - \mathbf{A}_\pi y(\rho, \theta)), \tag{39}$$

where $\pi = \pi_\theta$ for exact SVPO and $\pi = \pi_u$ for practical SVPO with u fixed.

Define $y = y(\rho, \theta)$ and $y' = y(\rho', \theta')$. Equation 43, together with the uniform moment bounds and continuity over the admissible policy family, yields a diagonal matrix $W \succ 0$ and $\lambda_{\mathrm{TD}} > 0$ such that, after the additive constant is fixed,

$$(y - y')^\top W \mathbf{A}_\pi (y - y') \geq \lambda_{\mathrm{TD}} \|y - y'\|^2 \tag{40}$$

for every admissible policy. For the next bound, take (ρ', θ') to be a critic equilibrium. Let π, π' be the exact policies at θ, θ', or the practical policies at $(u, \theta_H), (u, \theta'_H)$ for the same fixed u. The uniform bounds in Corollary 2 and Lemma 2, together with the Lipschitz estimates in Lemma 1, give $c_{\mathrm{TD}} < \infty$ such that

$$\|W\,[b_\pi - b_{\pi'} - (\mathbf{A}_\pi - \mathbf{A}_{\pi'})y']\| \leq c_{\mathrm{TD}} \|y - y'\|. \tag{41}$$

Accordingly, Assumption 7 requires $c_{\mathrm{TD}} < \lambda_{\mathrm{TD}}$ for convergence, with the common W and associated constants chosen uniformly over both policy families and all admissible parameters, including $u \in U$ for practical SVPO. An analogous condition appears in the SARSA convergence analysis of Melo et al. (2008).

C.12 Proof of Theorem 3

Proof. Let $\mathbf{A}_\pi^Q$ denote the lower-right block of $\mathbf{A}_\pi$ in equation 38. By stationarity of μ_π, for every vector v,

$$v^\top \mathbf{A}_\pi^Q v = \frac{1}{2} \iint (v^\top \xi(z) - v^\top \xi(z'))^2 \mu_\pi(dz) \mathsf{P}_\pi(z, dz'). \tag{42}$$

If the integral vanishes, then $v^\top \xi$ is constant along stationary trajectories. By ergodicity, $v^\top \xi(z)$ is therefore constant μ_π-almost everywhere. After fixing the additive-constant degree of freedom, Assumption 5 implies that $v^\top \xi(z)$ can be constant μ_π-almost everywhere only if $v = 0$. Hence

$$\iint (v^\top \xi(z) - v^\top \xi(z'))^2 \mu_\pi(dz) \mathsf{P}_\pi(z, dz') > 0 \qquad (v \neq 0). \tag{43}$$

Equation 43, compactness of the parameter sets, the common bound on $\mu_\pi(L^2)$, Assumption 7, and the invariant-measure continuity established in Lemma 1 give

$$\lambda_Q := \frac{1}{2} \inf_{\|v\|=1,\, \pi} \iint (v^\top \xi(z) - v^\top \xi(z'))^2 \mu_\pi(dz) \mathsf{P}_\pi(z, dz') > 0.$$

Here π ranges over the exact Gibbs policies for $\theta \in \Theta$ and the practical policies for $(u, \theta_H) \in U \times \Theta_H$.

The moment bound gives $\sup_\pi \|\int \xi(z) \mu_\pi(dz)\| =: M_\xi < \infty$. Choose $\kappa > M_\xi^2/(2\lambda_Q)$ and set $W = \mathrm{diag}(\kappa, I)$. Define $y - y' = (\rho - \rho', v)$. Young's inequality and equation 42 yield

$$\begin{aligned}(y - y')^\top W \mathbf{A}_\pi (y - y') &= \kappa(\rho - \rho')^2 + (\rho - \rho') v^\top \int \xi \, d\mu_\pi + v^\top \mathbf{A}_\pi^Q v \\ &\geq \left(\kappa - \frac{M_\xi^2}{2\lambda_Q}\right)(\rho - \rho')^2 + \frac{\lambda_Q}{2} \|v\|^2.\end{aligned}$$

Thus

$$\lambda_{\mathrm{TD}} := \min\left\{\kappa - \frac{M_\xi^2}{2\lambda_Q}, \frac{\lambda_Q}{2}\right\} > 0,$$

$$(y - y')^\top W \mathbf{A}_\pi (y - y') \geq \lambda_{\mathrm{TD}} \|y - y'\|^2,$$

which proves equation 40 uniformly over both policy families (Tsitsiklis & Van Roy, 1999).

Let (ρ^*, θ^*) be an interior critic equilibrium and let $y^* = y(\rho^*, \theta^*)$, with corresponding policy π_*. Since $\eta' > 0$, equation 39 implies

$$b_{\pi_*} - \mathbf{A}_{\pi_*} y^* = 0, \qquad \rho^* = J^{\pi_*}.$$

On the compact interval Θ_H, $\eta' > 0$ is bounded above and away from zero. After matching the additive constants, define

$$\mathcal{E}(\rho, \theta) := \frac{\kappa}{2}(\rho - \rho^*)^2 + \frac{1}{2}\|\theta_{\text{task}} - \theta^*_{\text{task}}\|^2 + \int_{\theta^*_H}^{\theta_H} \frac{\eta(s) - \eta(\theta^*_H)}{\eta'(s)}\, ds.$$

Then $\mathcal{E}$ is positive definite relative to y^*, and there exist $0 < c_1 \le c_2 < \infty$ such that

$$c_1\|y - y^*\|^2 \le \mathcal{E}(\rho, \theta) \le c_2\|y - y^*\|^2.$$

Along the unprojected critic ODE, equation 39, equation 40, and equation 41 give

$$\begin{aligned}
\dot{\mathcal{E}} &= (y - y^*)^\top W(b_\pi - \mathbf{A}_\pi y) \\
&= -(y - y^*)^\top W \mathbf{A}_\pi (y - y^*) + (y - y^*)^\top W\,[b_\pi - b_{\pi_*} - (\mathbf{A}_\pi - \mathbf{A}_{\pi_*})y^*] \\
&\le -(\lambda_{\text{TD}} - c_{\text{TD}})\|y - y^*\|^2.
\end{aligned}$$

Under the product projection in Assumption 6, Euclidean projection is applied separately to θ_{task} and θ_H. Since (ρ^*, θ^*) is an interior equilibrium and $\eta'(\theta_H) > 0$, projection cannot increase the directional derivative of $\mathcal{E}$. Hence

$$\begin{aligned}
\frac{d}{dt}\mathcal{E}(\rho(t), \theta(t))\Big|_{\text{projected field}} &\le \dot{\mathcal{E}} \\
&\le -(\lambda_{\text{TD}} - c_{\text{TD}})\|y - y^*\|^2.
\end{aligned}$$

Because $c_{\text{TD}} < \lambda_{\text{TD}}$, the equilibrium is globally attracting. Moreover, any zero of the projected critic field has $d\mathcal{E}/dt = 0$, and the strict decrease above then gives $y = y^*$. Thus the critic equilibrium is unique after the additive constant is fixed.

For practical SVPO, define $y^*(u) = y(\rho^*(u), \theta^*(u))$. Applying the preceding fixed-u decrease bound with $y = y^*(u')$, and using that $y^*(u')$ is an equilibrium at u' together with the u-Lipschitz bounds in Assumption 7, we have

$$(\lambda_{\text{TD}} - c_{\text{TD}})\|y^*(u) - y^*(u')\| \le C\|u - u'\|.$$

Thus $(\rho^*(u), \theta^*(u))$ is Lipschitz in u.

The compact projections bound u_t and θ_t. The first coordinate of the Markov-noise conclusion in Lemma 2 shows that

$$S_n := \sum_{t=0}^{n-1} \beta_t (c(X_t, A_t) - J^{\pi_t})$$

converges almost surely. The cost growth condition in Assumption 5 and the common invariant moment bound give $\sup_t |J^{\pi_t}| < \infty$. With $\widetilde{\rho}_n = \rho_n - S_n$, the scalar recursion becomes

$$\widetilde{\rho}_{n+1} = (1 - \beta_n)\widetilde{\rho}_n + \beta_n(J^{\pi_n} - S_n).$$

Choose t_0 such that $0 < \beta_n \le 1$ for every $n \ge t_0$. Since S_n converges almost surely,

$$\begin{aligned}
\sup_{n \ge t_0} |\widetilde{\rho}_n| &\le \max\left\{ |\widetilde{\rho}_{t_0}|, \sup_{n \ge t_0} |J^{\pi_n} - S_n| \right\} < \infty, \\
\sup_{n \ge t_0} |\rho_n| &\le \sup_{n \ge t_0} |\widetilde{\rho}_n| + \sup_{n \ge t_0} |S_n| < \infty \qquad \text{a.s.}
\end{aligned}$$

The finitely many iterates before t_0 are also finite almost surely, so $\sup_n |\rho_n| < \infty$ almost surely.

Set $\widetilde{Z}_{t+1} = (X_{t+1}, \widetilde{A}_{t+1})$. The sampled critic noise decomposes as

$$H_c(Z_t, \widetilde{Z}_{t+1}; \rho_t, \theta_t) - \bar{h}_c(u_t, \rho_t, \theta_t)$$

$$= \Big[H_c(Z_t, \widetilde{Z}_{t+1}; \rho_t, \theta_t) - h_c(Z_t; u_t, \rho_t, \theta_t)\Big] + \big[h_c(Z_t; u_t, \rho_t, \theta_t) - \bar{h}_c(u_t, \rho_t, \theta_t)\big] .$$

According to the definition of $\mathcal{F}_t$, the first bracket term is then a martingale difference under the pre-update policy. By equation 35,

$$\mathsf{E}\sum_t \beta_t^2 \mathsf{E}\left[\left\|H_c(Z_t, \widetilde{Z}_{t+1}; \rho_t, \theta_t) - h_c(Z_t; u_t, \rho_t, \theta_t)\right\|^2 \,\middle|\, \mathcal{F}_t\right]$$
$$\leq C\sum_t \beta_t^2 \mathsf{E}\left[1 + L^{2m}(X_t) + |\rho_t|^2(1 + L^2(X_t))\right] < \infty.$$

Hence the corresponding weighted martingale-difference series converges almost surely. The second bracket is controlled by Lemma 2.

The boundedness and Lipschitz estimates above, the two noise bounds, the step-size conditions in Assumption 6, and the global projected-field decrease therefore satisfy the hypotheses of the standard two-timescale ODE theorem (Borkar, 2008, Chapter 6). For practical SVPO, after matching the relative-value additive constants, it follows that

$$\|(\rho_t, \theta_t) - (\rho^*(u_t), \theta^*(u_t))\| \longrightarrow 0 \qquad \text{a.s.}$$

Consequently

$$\|\bar{h}_c(u_t, \rho_t, \theta_t)\| \longrightarrow 0 \qquad \text{a.s.}$$

For exact SVPO, the same projected-ODE and noise argument applies directly, yielding the corresponding critic-residual convergence under π_{θ_t}.

The additive-constant direction fixed in the preceding critic analysis does not affect the convergence conclusion. Its mean TD update equals the average-cost residual, which is the first component of $\bar{h}_c$, and hence also vanishes as $\|\bar{h}_c\| \to 0$. Moreover, a global additive constant cancels from the Gibbs policy and contributes zero to the invariant actor mean by the score identity. Thus the preceding conclusions extend to the full critic field.

Now suppose $\bar{h}_a(u, \theta^*(u)) = \nabla F(u)$. Critic tracking and Lemma 1 imply

$$\|\bar{h}_a(u_t, \theta_t) - \nabla F(u_t)\| \longrightarrow 0.$$

The actor part of Lemma 2, together with the two-timescale separation, implies that the actor iterates track the projected gradient flow generated by $-\nabla F$. Along every solution of this flow,

$$F(u(T)) - F(u(0)) = -\int_0^T \|\dot{u}(t)\|^2 \, dt \leq 0.$$

The invariance principle therefore gives

$$\|u_t - \Gamma_U[u_t - \nabla F(u_t)]\| \to 0.$$

By nonexpansiveness of Γ_U,

$$\|u_t - \Gamma_U[u_t - \bar{h}_a(u_t, \theta_t)]\|$$
$$\leq \|u_t - \Gamma_U[u_t - \nabla F(u_t)]\| + \|\bar{h}_a(u_t, \theta_t) - \nabla F(u_t)\| \to 0.$$

If the projection is inactive at the limit points, projected stationarity reduces to $\nabla F = 0$ there. Compactness of U and continuity of ∇F then imply $\|\nabla F(u_t)\| \to 0$. □

C.13 Global-optimality regime

Let J^* denote the globally minimal average cost among stationary policies with a well-defined invariant average cost. Suppose $\bar{h}_a(u, \theta^*(u)) = \nabla F(u)$ and define $F^* := \min_{u \in U} F(u)$. Assume there exists $C_{\mathrm{PL}} < \infty$ such that

$$F(u) - F^* \leq C_{\mathrm{PL}} \|u - \Gamma_U[u - \nabla F(u)]\|^2. \tag{44}$$

Assume further that $F(u) = J^{\pi_u}$ at the critic equilibrium $(\rho^*(u), \theta^*(u))$ and that this policy family contains a globally optimal stable policy.

Corollary 5 (Global optimality)**.** *Under the hypotheses of Theorem 3 and the conditions above, practical SVPO satisfies $F(u_t) \to F^*$ and $J^{\pi_t} \to J^*$ almost surely.*

C.14 Proof of Corollary 5

Proof. Theorem 3 and critic tracking give

$$\|u_t - \Gamma_U[u_t - \nabla F(u_t)]\| \to 0 \qquad \text{a.s.}$$

Equation 44 therefore yields

$$0 \le F(u_t) - F^* \le C_{\mathrm{PL}} \|u_t - \Gamma_U[u_t - \nabla F(u_t)]\|^2 \to 0.$$

Since $F(u) = J^{\pi_u}$ at the critic equilibrium and the policy family contains a globally optimal stable policy,

$$J^* \le F^* = \min_{u \in U} F(u) \le J^*.$$

Hence $F^* = J^*$.

At the critic equilibrium, the first component of equation 13 gives $\rho^*(u) = J^{\pi_u} = F(u)$. Therefore, the critic-residual and tracking conclusions of Theorem 3 imply

$$|J^{\pi_t} - F(u_t)| \le \|\bar{h}_c(u_t, \rho_t, \theta_t)\| + |\rho_t - \rho^*(u_t)| \to 0 \qquad \text{a.s.}$$

Finally,

$$|J^{\pi_t} - J^*| \le |J^{\pi_t} - F(u_t)| + |F(u_t) - F^*| \to 0 \qquad \text{a.s.}$$

□

C.15 Proof of Corollary 3

Proof. Fix $1 \le m \le \widetilde{m}$ and let $\mathcal{V}(x) = 1 + L^m(x)$. By assumption, $\|Q^\pi - \widehat{Q}_{\theta_t}\|_{\mathcal{V}} < \infty$. By Gibbs optimality and the realization-gap bound,

$$\begin{aligned} \mathcal{G}_x(\pi_t; \theta_t) &\le \mathcal{G}_x(\pi; \theta_t) + \bar{\epsilon}_\pi (1 + L(x)^{q/k}) \\ &\le \mathcal{G}_x(\pi; \theta_t) + \bar{\epsilon}_\pi \mathcal{V}(x), \end{aligned}$$

where the last inequality uses $L \ge 1$ and $q/k < m$. Expanding the free energies and using nonnegativity of KL gives

$$\int \widehat{Q}_{\theta_t} \, d(\pi_t - \pi) \le \bar{\epsilon}_\pi \mathcal{V}(x) + \tau \mathrm{KL}(\pi(\cdot \mid x) \| \nu).$$

Adding and subtracting $\widehat{Q}_{\theta_t}$ then yields

$$\begin{aligned} &\int Q^\pi(x,a) \pi_t(da \mid x) - \int Q^\pi(x,a) \pi(da \mid x) \\ &= \int (Q^\pi - \widehat{Q}_{\theta_t}) \, d\pi_t + \int \widehat{Q}_{\theta_t} \, d(\pi_t - \pi) + \int (\widehat{Q}_{\theta_t} - Q^\pi) \, d\pi \\ &\le \left(2\|Q^\pi - \widehat{Q}_{\theta_t}\|_{\mathcal{V}} + \bar{\epsilon}_\pi \right) \mathcal{V}(x) + \tau \sup_y \mathrm{KL}(\pi(\cdot \mid y) \| \nu). \end{aligned}$$

By the comparator Poisson equation,

$$\begin{aligned} &\int Q^\pi(x,a) \pi(da \mid x) = V^\pi(x), \\ &Q^\pi(x,a) - V^\pi(x) = c(x,a) - J^\pi + \mathsf{P} V^\pi(x,a) - V^\pi(x). \end{aligned}$$

Therefore, invariance of μ_{π_t} gives

$$\begin{aligned} \iint (Q^\pi(x,a) - V^\pi(x)) \mu_{\pi_t}(dx, da) &= \mu_{\pi_t} (c - J^\pi + \mathsf{P} V^\pi - V^\pi) \\ &= J^{\pi_t} - J^\pi. \end{aligned}$$

Integrating the preceding pointwise inequality with respect to μ_{π_t} thus yields

$$J^{\pi_t} - J^{\pi} \le \left(2\|Q^{\pi} - \widehat{Q}_{\theta_t}\|_{\mathcal{V}} + \bar{\epsilon}_{\pi}\right) \mu_{\pi_t}(\mathcal{V}) + \tau \sup_{y} \mathrm{KL}(\pi(\cdot \mid y)\|\nu).$$

Finally, Theorem 2 gives CPPI(∞) for π_t, and hence CPPI(m). The invariant-moment bound in Theorem 1 therefore gives

$$\mu_{\pi_t}(\mathcal{V}) = 1 + \mu_{\pi_t}(L^m) \le 1 + \frac{\widetilde{b}_{m,\pi_t}}{1 - \widetilde{\lambda}_{m,\pi_t}}.$$

Substituting this bound proves equation 15. □

D Experimental Environments and Extended Results

We give the environments, protocols, and extended results for Section 5. Comparisons use ten matched seeds with identical environment settings, physical costs, task features, base task-component initialization and training budgets. Certificate statistics evaluate the learned *raw* policy. LTI runs last 50,000 steps with checkpoints every 100 steps; LQR and queueing runs last 200,000 steps with checkpoints every 1,000 steps.

D.1 Metrics and update-direction diagnostic

We record raw-policy tail margins, average task cost $\bar{J}_t = t^{-1}\sum_{s=0}^{t-1} c(X_s, A_s)$ for $t \ge 1$, tail occupancy $P_{\mathrm{tail}}(t) = \Pr[L(X) > R_{\mathrm{cal}}]$, TD-error variance, and physical escape. These are finite-grid and finite-radius diagnostics. A persistent crossing consists of three consecutive non-positive margin evaluations. In all three settings, $\widehat{\mathcal{M}}_t$ is the minimum normalized raw-policy drift margin over the fixed tail-state set, computed separately for each seed. Recent cost, tail occupancy, and TD variance use the latest 100 training observations for LTI and 1,000 for LQR. LQR task costs average observations through escape or the full horizon; queueing task costs summarize completed runs.

We compare two updates from the same frozen policy: $h^*(u)$ uses accurate task values, and $\widehat{h}(u)$ uses approximate task values. Both follow the same improvement rule. At differentiable points of the margin, define

$$D^*(u) = \nabla_u \widehat{\mathcal{M}}(u)^{\top} h^*(u), \qquad \widehat{D}(u) = \nabla_u \widehat{\mathcal{M}}(u)^{\top} \widehat{h}(u).$$

The sign pattern $D^* > 0$ and $\widehat{D} < 0$ means that ideal task improvement increases the margin while the practical update decreases it. We also compare the first persistent crossing time $T_{\widehat{\mathcal{M}}}$ with the onset of persistent tail growth and TD variance, denoted T_{tail} and T_{TD}. These quantities are offline diagnostics for the mechanism analysis.

D.2 Switched-LTI mechanism audit

The scalar system is

$$X_{t+1} = 1.1X_t + U_t + W_t = f_{A_t} X_t + W_t, \qquad W_t \sim \mathcal{N}(0, 0.08^2),$$

where action a applies $U_t = K_a X_t$ and

$$f_a = 1.1 + K_a \in \{0.2, 0.3, 0.45, 0.6, 0.8, 1.0, 1.15, 1.25, 1.35\}.$$

The Gibbs base measure ν is uniform on these nine actions. The physical task cost is $c(x,a) = (1 + 0.2K_a^2)x^2$, the task feature is $e_a \otimes [1, x^2]$, and the Lyapunov function is $L(x) = 1 + x^4$, with certificate feature $\phi_H(x,a) = |f_a x|^4$. Thus the task and certificate orders are $q = 2$ and $k = 4$. The exact one-step certificate expectation is

$$\mathbb{E}[L(X^+) \mid x, a] = 1 + f_a^4 x^4 + 6 f_a^2 (0.08)^2 x^2 + 3(0.08)^4.$$

LTI variants share the task initialization, with $\tau = 0.2$, $X_0 \sim \mathcal{N}(0, 0.2^2)$, and $\rho_0 = 0$. SVPO uses $\theta_{H,0} = -1$ and certificate floor 0.05. Critic and practical-actor projections have radius 30. At

Table 1: Switched-LTI persistent raw-policy crossings and physical escapes over ten matched seeds.

λ	Method	Crossing	Escape
0	Vanilla AC	9/10	0/10
0	SVPO	0/10	0/10
0.2	Vanilla AC	10/10	8/10
0.2	SVPO	0/10	0/10
0.5	Vanilla AC	10/10	10/10
0.5	SVPO	0/10	0/10

update $n \geq 1$, learning rates follow $\ell_0(1 + \max\{n - 100, 0\}/500)^{-p}$, with $(\ell_0, p) = (0.05, 0.6)$ for the critic, $(0.02, 0.6)$ for average cost, and $(0.02, 0.8)$ for the practical actor. For LTI, we use the tail-state grid $\mathcal{G}_{\text{tail}} = \{4, 6, 8, 12, 16\}$ for margin and excess-load evaluations. The prespecified grid lies outside the compact set $|x| \leq 3.1$ used in the reference-policy drift bound and is shared across methods and bias strengths. Positive states suffice because both diagnostics are even in x. Let $f_* = \min_a |f_a|$, $f_{\max} = \max_a |f_a|$, and $\sigma = 0.08$. The true action contrast and excess dissipativity load are

$$\chi_L(x, a) = (f_a^4 - f_*^4)x^4 + 6\sigma^2(f_a^2 - f_*^2)x^2,$$
$$s(x, a) = (f_a^4 - f_*^4)x^4.$$

In particular, $0 \leq \chi_L(x, a) - s(x, a) \leq 6\sigma^2(f_{\max}^2 - f_*^2)x^2 \leq C_R(1 + L(x)^{1/2})$. Hence equation 16 holds with $\underline{c} = \overline{c} = 1$ and $q/k = 1/2$.

The same finite-action model also verifies the two mass hypotheses used by the certificate-floor theorem. The action with coefficient $f_* = 0.2$ has $s = 0$, so Assumption 3 holds with $\xi_0 = 0$ and $p_0 = 1/9$. Let $d = f_{\max}^4 - f_*^4 > 0$, $x_n = n \geq 1$, and $R_n = \{a : |f_a| = f_{\max}\}$. Then $\nu(R_n) = 1/9$ and

$$\frac{d}{2} L(x_n) \leq d x_n^4 = s(x_n, a) \leq d L(x_n) \qquad (a \in R_n).$$

Thus equation 28 holds with $(c_-, c_+, p_1) = (d/2, d, 1/9)$. The switched-LTI model therefore satisfies the hypotheses of Proposition 1 and Corollary 4; the necessary floor in this instance is $\eta \geq \tau\gamma/2$. This evaluator also admits a theorem-aligned drift decomposition. Define

$$r_\sigma(x) := \max_a \{6 f_a^2 \sigma^2 x^2 + 3\sigma^4\},$$
$$\mathcal{D}_0(x) := x^4 - r_\sigma(x), \qquad \mathcal{D}(x) := [\mathcal{D}_0(x)]_+.$$

Take $\mathcal{C} = \{x : L(x) \leq R\}$ so that $\mathcal{C}$ contains $\{x : \mathcal{D}_0(x) < 0\}$ and R exceeds the reference-policy tail threshold. With the compact remainder absorbed into finite b_0, b_1, Assumption 1 is satisfied:

$$\mathbb{E}[L(X^+) \mid x, a] - L(x) \leq \phi_H(x, a) - \mathcal{D}(x) + b_1 \mathbb{I}_{\mathcal{C}}(x).$$

The deterministic reference action $f = 0.8$ obeys the reference-policy drift outside $\mathcal{C}$ for any fixed $c_0 \in (0, 1 - 0.8^4)$. Thus the finite-cost task has an explicit stable witness. For every finite m,

$$\frac{\mathbb{E}[(1 + (f_a x + W)^4)^m]}{(\mathbb{E}[1 + (f_a x + W)^4])^m} \longrightarrow 1 \qquad \text{as } |x| \to \infty$$

uniformly over the nine actions, and all corresponding moments are finite on compact sets. This verifies Assumption 2. The positive Gaussian density gives irreducibility, aperiodicity, and compact-set petiteness. On each compact sublevel, the finite action set and compact projected SVPO parameter sets also give a common lower bound on the Gibbs action probabilities, which supplies Assumption 8 for the state–action kernels.

The controlled low-order actor bias is active only during steps $[6000, 14000)$ and subtracts λx^2 from the raw sampling energy of the $f = 1.25$ action. It changes the raw sampling energy while holding the system, cost, critic target, and Lyapunov function fixed. We record a physical escape and terminate the LTI run at the first step with $|X_t| \geq 100$.

Table 2: Certificate-pathway audit with the same task critic and no actor bias. Minimum raw margin is the median across ten seeds ($\lambda = 0$).

Method	Crossing	$\min_t \widehat{\mathcal{M}}_t$
SVPO (Gibbs)	0/10	0.99388
CriticOnly	10/10	-2.32180
No-ϕ_H	10/10	-1.44882

Table 1 separates certificate crossing from physical escape. At $\lambda = 0$, Vanilla AC crosses in $9/10$ seeds with $0/10$ escapes. At $\lambda = 0.2$ and 0.5, every Vanilla AC seed crosses, while the escape rate rises to $8/10$ and $10/10$. Certificate crossing is an earlier and more sensitive failure indicator than physical escape, while stronger task pressure makes physical escape increasingly likely after crossing. SVPO shows no persistent margin crossings or physical escapes at any of the three pressure levels. For the escaping $\lambda = 0.5$ trajectories, the persistent margin crossing precedes physical escape; the corresponding learning and state-tail trajectories appear in Figures 1 and 5. After a run terminates upon reaching the physical escape threshold, its final observation is carried forward only for visualization.

Certificate pathway. Table 2 compares three learners with exact Gibbs policies with the same quadratic task critic and zero actor bias to isolate the role of the certificate in the sampling policy. SVPO and CriticOnly share the certificate-augmented critic and TD update, but CriticOnly omits the certificate from the sampling policy; No-ϕ_H omits it from both critic and policy. SVPO has no certificate crossings in any of the ten seeds, whereas CriticOnly and No-ϕ_H cross in all ten. Thus, including the certificate in the critic alone is insufficient; transferring it to the sampling policy is what preserves the raw-policy certificate. All three learners complete $50{,}000$ steps without physical escape, isolating this effect before escape.

Graded excess-load moments. To examine the higher-grade conditions in CPPI, we evaluate the normalized conditional excess-load moments of the practical actor. For each seed ℓ and moment order j, define

$$R_j^{(\ell)} = \max_{t \in \mathcal{T}_\ell} \max_{x \in \mathcal{G}_{\mathrm{tail}}} \frac{\mathbb{E}_{\pi_t}[s(x,A)^j \mid x]}{1 + L(x)^{j/2}}, \qquad j = 1, \dots, 4.$$

Here $\mathcal{T}_\ell$ contains the observed regular checkpoints for seed ℓ, including initialization. Figure 4 shows a clear separation across all four moment orders: the normalized loads of Vanilla AC grow rapidly with j, whereas those of SVPO remain several orders of magnitude smaller. This pattern is consistent across all three tested bias strengths, showing that SVPO controls not only the grade-one margin but also the higher-order excess-load moments appearing in CPPI.

D.3 Input-saturated LQR

We use the two-dimensional system

$$x_{t+1} = Ax_t + Bu_t + w_t, \qquad u_t \in [-1, 1], \qquad A = \mathrm{diag}(1, 0.9), \quad B = (0.25, 0)^\top,$$

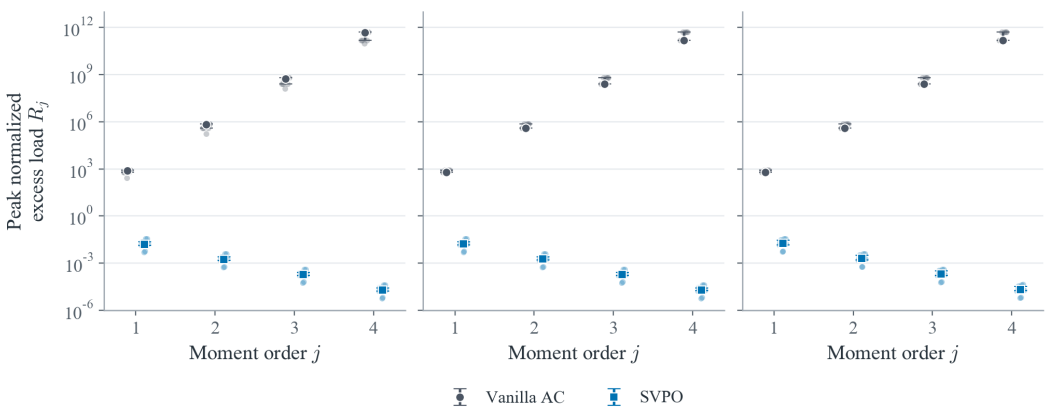
with Gaussian noise covariance $\Sigma_W = \mathrm{diag}(0.06^2, 0.03^2)$ and physical cost $c(x,u) = x^\top Q x + 0.1u^2$ for $Q = \mathrm{diag}(1, 0.1)$. Every method uses the action-aware actor feature

$$\psi_{\mathrm{act}}(x,u) = [1, x_1^2, x_1x_2, x_2^2, u^2, \bar{x}_1^+, \bar{x}_2^+, (\bar{x}_1^+)^2, \bar{x}_1^+\bar{x}_2^+]^\top, \qquad \bar{x}^+ = Ax + Bu,$$

at temperature $\tau = 0.15$. The shared task critic uses $[1, x_1^2, x_1x_2, x_2^2, u^2]^\top$, preserving the immediate action cost but omitting the nominal-next-state terms. The certificate is

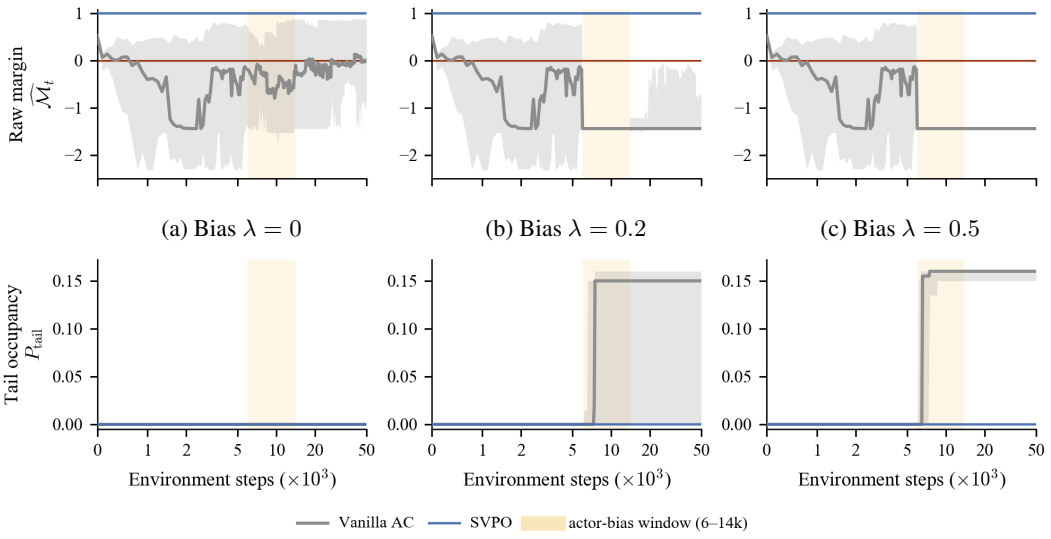
$$L(x) = 1 + (x^\top P x)^2, \qquad \phi_H(x,u) = ((Ax+Bu)^\top P (Ax+Bu))^2, \qquad P = \mathrm{diag}(1, 0.25).$$

This LQR experiment tests the certificate-preserving raw-policy mechanism under continuous actions and critic approximation, beyond the sufficient conditions of our theory. For the chosen quartic L, bounded inputs give $\mathsf{P}_\pi L(r,0)/L(r,0) \to 1$ as $r \to \infty$, so the geometric reference-policy drift required by Assumption 1 does not hold. Moreover, under the uniform action base,



(a) Bias $\lambda = 0$ (b) Bias $\lambda = 0.2$ (c) Bias $\lambda = 0.5$

Figure 4: Peak finite-grid excess-load ratios $R_j^{(\ell)}$ for the practical actor. Small dots show ten seeds; large markers and error bars show medians and interquartile ranges. All panels share a logarithmic vertical scale.



(a) Bias $\lambda = 0$ (b) Bias $\lambda = 0.2$ (c) Bias $\lambda = 0.5$

(d) Bias $\lambda = 0$ (e) Bias $\lambda = 0.2$ (f) Bias $\lambda = 0.5$

Figure 5: **Switched-LTI pressure response.** Columns correspond to $\lambda \in \{0, 0.2, 0.5\}$; rows show raw-policy margin and state-tail occupancy. Curves are medians with 10th–90th percentile ranges over ten matched seeds. The horizontal axis is linear through 10^3 steps and logarithmic thereafter. Shading marks the prespecified 6,000–14,000-step intervention window, with zero amplitude in the $\lambda = 0$ control. Vanilla's zero-pressure crossings do not reach the state tail, whereas the two positive pressures produce tail occupancy; SVPO preserves a positive margin and suppresses tail visits throughout. Terminated Vanilla trajectories retain their final observation only for visualization.

$\nu\{u : s((r,0),u) \le \xi_0\} = O(r^{-3})$, violating the uniform action-coverage constant in Assumption 3. Thus, the theoretical recurrence and mixing guarantees are not invoked for this experiment. Instead, we directly evaluate the learned raw-policy margin to test whether the certificate-preserving mechanism remains effective in this more demanding setting.

The reported runs use seeds 0–9, 200,000 training steps, and evaluations every 1,000 steps. Initial critic, average-cost, and actor step sizes are respectively 0.05, 0.02, and 0.02; after a 100-step hold

Table 3: **Critic approximation reverses the LQR margin response.** Positive values indicate an increasing margin; negative values indicate a decreasing margin.

Evaluation grid (x_1 range, points)	Accurate values D^*_ϵ	Restricted critic $\widehat{D}_\epsilon$
$[-12, 12]$, 201	$+0.0607$	-0.0562
$[-12, 12]$, 301	$+0.0600$	-0.0562
$[-16, 16]$, 301	$+0.0605$	-0.0562
$[-16, 16]$, 401	$+0.0600$	-0.0562

they decay on a 10,000-step scale with powers 0.6, 0.6, and 0.8. The critic and actor projection radii are 800 and 80, respectively. The tail radius is $\|x\|_P = 4$, and the physical escape radius is $\|x\|_P = 5.7$. The frozen tail audit uses radii $\{2, 3, 4, 6, 8, 12\}$ and eight angles at each radius. Policy integrals use 32-point Gauss–Legendre quadrature, and sampling uses a 65-point action grid. The diagnostic P_{good} averages $\pi(\text{sign}(x_1)u \le -0.5 \mid x)$ uniformly over tail-grid states satisfying $|x_1| \ge 0.25\|x\|_P$.

Effect of critic approximation. To isolate the effect of critic approximation on the policy-improvement direction, we start from the same initial Vanilla policy and keep the actor rule fixed. We audit the controlled coordinate x_1 on the slice $x_2 = 0$, since the second state coordinate is stable and unaffected by the control input. Accurate task values are computed from the centered average-cost Poisson system on the grids in Table 3, using Gaussian quadrature, linear interpolation, and endpoint clipping. The restricted critic solves the population expected-SARSA equations by least squares. Both directions use the same stationary state–action weights. Let $h^*(u)$ and $\widehat{h}(u)$ denote the update directions obtained from accurate long-term task values and from the restricted critic, respectively. We define the corresponding normalized update directions and estimate their normalized responses using $\epsilon = 10^{-3}$:

$$D^*_\epsilon(u) = \frac{\widehat{\mathcal{M}}(u + \epsilon h^*(u)/\|h^*(u)\|) - \widehat{\mathcal{M}}(u)}{\epsilon},$$
$$\widehat{D}_\epsilon(u) = \frac{\widehat{\mathcal{M}}(u + \epsilon \widehat{h}(u)/\|\widehat{h}(u)\|) - \widehat{\mathcal{M}}(u)}{\epsilon}.$$

Table 3 shows the same sign reversal across all four evaluation grids: $D^*_\epsilon > 0$, whereas $\widehat{D}_\epsilon < 0$. Thus, in this controlled comparison, critic approximation reverses the effect of the policy update on the raw-policy margin.

Methods and performance. Vanilla+Tilt and Lagrangian AC use the same ϕ_H as SVPO; Filter AC uses the same L and known one-step dynamics. Vanilla+Tilt applies equation 23 with fixed coefficient 0.01. Entropy AC uses an entropy coefficient of 0.15 in both the TD target and the actor update.

For Lagrangian AC, the state-dependent certificate budget is

$$\left[\|x\|_P^4 - \max_{u \in [-1,1]} \{\mathsf{P}L(x,u) - 1 - \phi_H(x,u)\}\right]_+ .$$

The actor energy is augmented by $\mu_t \phi_H$, and the TD target includes the corresponding certificate-budget residual weighted by μ_t. The multiplier is initialized at $\mu_0 = 0$, projected onto $[0, 20]$, and updated with an initial step size 10^{-3} using the actor's decay schedule.

Filter AC numerically searches for the action closest to the sampled raw action subject to

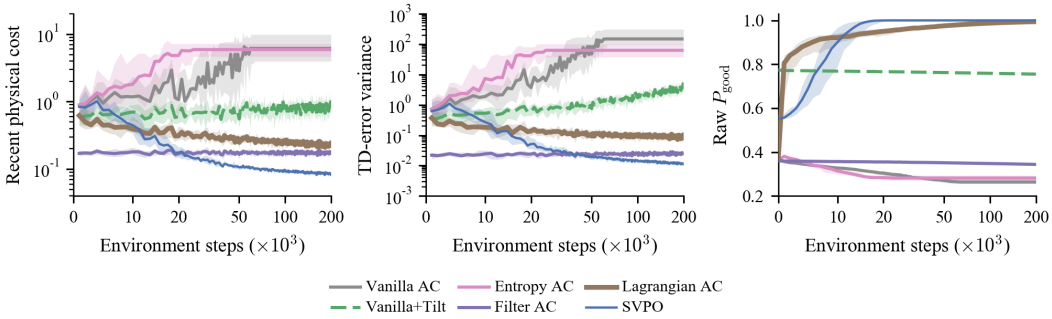
$$\mathsf{P}L(x,u) \le L(x) + 0.05(1 - \|x\|_P^4).$$

If no feasible action is found, it selects the action minimizing $\mathsf{P}L(x,u)$. Raw and executed policies are evaluated separately.

SVPO uses $\eta(\theta_H) = \text{Softplus}(\theta_H) + 10^{-6}$, initialized to $\eta_0 = 10^{-3}$. The base task-actor and task-critic parameters match those of Vanilla, giving initial raw margins of 0.01883696 for SVPO

Table 4: Input-saturated LQR results over ten matched seeds. Crossing and the minimum margin use the raw learned policy. Escape uses the executed process (and hence the filtered process for Filter AC). Cost is the median across seeds of each run's average physical cost up to escape or the 200,000-step horizon.

Method	Crossing	Escape	$\min_t \widehat{\mathcal{M}}_t$	Cost
Vanilla AC	10/10	10/10	-0.02117	2.03479
Vanilla+Tilt	0/10	0/10	0.05160	0.78552
Entropy AC	10/10	10/10	-0.02758	2.28301
Lagrangian AC	0/10	0/10	0.01202	0.27127
Filter AC	0/10	0/10	0.00698	0.17552
SVPO	0/10	0/10	0.01816	0.13828



(a) Recent physical cost (b) TD variability (c) Raw inward-action mass

Figure 6: **Additional input-saturated LQR diagnostics.** The panels show recent physical cost, TD-error variance, and the raw-policy inward-action probability P_{good}. For Filter AC, physical cost uses executed actions and P_{good} evaluates the unfiltered raw policy. Curves are medians with 10th–90th percentile ranges over ten matched seeds. The horizontal axis is linear through 10^4 steps and logarithmic thereafter. Terminated Vanilla AC and Entropy AC trajectories retain their final observation only for visualization.

and 0.01202435 for Vanilla. SVPO follows the TD update in Algorithm 1, retains the physical task cost, and shares the critic's certificate channel with the raw policy.

Table 4 reports crossings, escapes, margins, and costs. Among the methods that complete training, SVPO attains the lowest full-process cost of 0.13828, 49.0% lower than Lagrangian AC and 21.2% lower than Filter AC, with zero crossings and escapes in all ten seeds. For Filter AC, the median minimum executed-policy margin is 0.05671, with a median filter intervention rate of 16.43%. To rule out initialization effects, we also test an SVPO variant with the same initial raw policy as Vanilla. It completes all ten runs without margin crossings or escapes.

Figure 6 further separates raw-policy certificate preservation from task learning under critic approximation. Vanilla+Tilt maintains positive raw-policy margins and completes training without escape, while its task cost and TD-error variance increase during training. This is consistent with Tilt's role as a portable KL-minimal certificate correction: it improves certificate behavior while leaving the underlying learner unchanged.

SVPO instead incorporates certificate preservation directly into the learning recursion while retaining the original physical task cost. In this experiment, it maintains positive raw-policy margins with lower TD-error variability and achieves the lowest full-process task cost among the methods that complete training. The comparison thus highlights the complementary roles of the two mechanisms: Tilt transfers certificate preservation to an existing learner with minimal intervention, whereas SVPO integrates it with continued learning and task optimization.

D.4 UNBOUNDED QUEUEING SYSTEM

We consider a two-queue system in which each action prioritizes one queue while the other experiences weaker service. The queue state $q_t \in \mathbb{R}_+^2$ evolves as

$$q_{t+1} = F_{A_t} q_t + A_t^{\text{arr}}, \qquad F_1 = \text{diag}(0.8, 1.15), \quad F_2 = \text{diag}(1.15, 0.8),$$

where the two arrival coordinates are independent $\text{Unif}[0.05, 0.15]$ variables. Selecting action a contracts the served queue and expands the other. The nonnegative energy–delay cost is

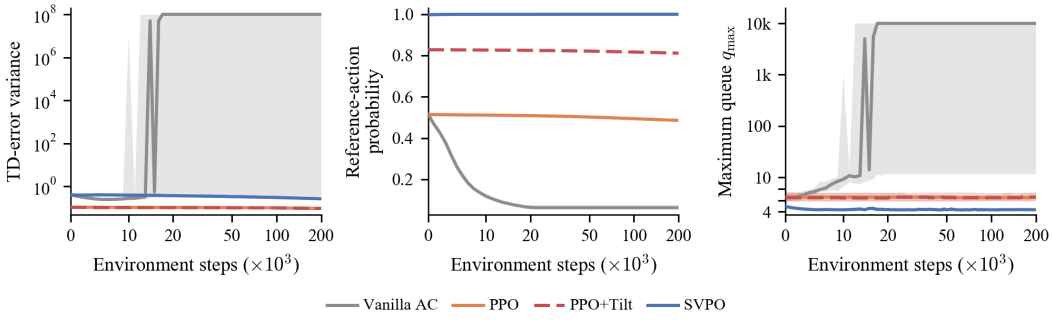
$$c_t = \|q_{t+1}\|_1 + 5\frac{q_{t,a_t}}{2 + q_{t,a_t}},$$

and the task features are $e_a \otimes [1, \sqrt{q_1}, \sqrt{q_2}, q_1, q_2]$. We use $L(q) = 1 + \|q\|_2^2$ and $\phi_H(q, a) = \|F_a q\|_2^2$. Uniform randomization is a stable policy witness because $(0.8^2 + 1.15^2)/2 = 0.98125 < 1$. The reference actions $\arg\min_a \phi_H(q, a)$ serve a largest queue, including both actions at ties. Figure 7 reports their total raw-policy probability, averaged equally over the fixed tail-state set. All methods initialize $q_{0,i} \sim \text{Unif}[0, 0.5]$ independently and share initial task energy $0.55(q_a - 2\sqrt{11 q_a})$. The fixed set $\mathcal{G}_{\text{tail}}$ contains 131 balanced and imbalanced workload states. For $\widehat{\mathcal{M}}_t$, the one-step expectation is evaluated exactly using $\mathbb{E}[L(q^+) \mid q, a] = 1 + \|F_a q\|_2^2 + 2(F_a q)^\top \mathbb{E}[A_t^{\text{arr}}] + \mathbb{E}[\|A_t^{\text{arr}}\|_2^2]$. We average this quantity over actions under $\pi_t(\cdot \mid q)$, form the normalized margin at each state, and take its minimum over $\mathcal{G}_{\text{tail}}$. The threshold $R_{\text{cal}} \approx 14.53$ is the 95th percentile of $L(q)$ over 50,000 steps under the larger-queue-first policy. Escape is recorded at the first of five consecutive steps with $L(q) > 10^8$, or immediately at $L(q) > 10^{10}$.

At each checkpoint, frozen-policy evaluations use 5,000 warmup and $N = 5{,}000$ measurement steps with matched randomness. For the measured successor states q_j^+, tail occupancy and maximum queue length are

$$P_{\text{tail}}(t) = \frac{1}{N}\sum_{j=1}^{N} \mathbf{1}\{L(q_j^+) > R_{\text{cal}}\}, \qquad q_{\max}(t) = \max_{j,i} q_{j,i}^+.$$

Evaluations stop at $L(q) \geq 10^8$, with fixed penalties for the remaining horizon and a lower bound on reported TD variance.



(a) TD variability (b) Reference-action selection (c) Maximum queue

Figure 7: **Additional queueing diagnostics.** The panels show TD-error variance, the tail-grid mean probability of serving a largest queue, and maximum queue length. Curves are medians with 10th–90th percentile ranges over ten matched seeds. Curves start from evaluations of each initialized policy. The horizontal axis is linear through 10^4 steps and logarithmic thereafter. Vanilla AC's terminal diagnostic is retained after its registered escape only for visualization; Table 6 reports the corresponding terminal quantities as censored.

Vanilla AC and SVPO use Gibbs actors at $\tau = 1$ and one SARSA update per transition. Critic and average-cost updates use $\beta_t = 10^{-3}(1 + t/5000)^{-0.8}$, with $\rho_0 = 0$. Task parameters have projection radius 50, with $\theta_H \in [-50, 50]$. SVPO initializes its certificate coefficient 0.2 above the PPO+Tilt coefficient.

PPO implementation. PPO uses a linear Gibbs actor at temperature $\tau = 1$ on the task features above and a separate linear differential reward critic $\widehat{V}$ on $[1, \sqrt{q_1}, \sqrt{q_2}, q_1, q_2]$. For reward $-c_t$, the average-reward TD advantage is

$$\delta_t = -c_t + \rho_t + \widehat{V}(q_{t+1}) - \widehat{V}(q_t),$$

where ρ_t estimates average physical cost. Actor advantages are standardized within each batch; the critic fits targets $\widehat{V}(q_t) + \delta_t$ by squared error. The average-cost estimate starts from the first batch mean and is subsequently updated toward each batch mean with weight approximately 0.6323.

PPO and PPO+Tilt use minibatch stochastic gradient updates with the settings in Table 5. Actor and critic learning rates decay polynomially with exponents 0.8 and 0.6, respectively, on a 10^5-step scale. PPO+Tilt starts from the same task-actor parameters as PPO and uses energy $w_u(q, a) + \eta\phi_H(q, a)$ with fixed $\eta = 10^{-2.25} \approx 0.0056234$. For coefficient calibration before training, we select the smallest value on $\{10^{-9+j/4} : j = 0, \ldots, 36\}$ whose initial median of statewise normalized drift margins over $\mathcal{G}_{\mathrm{tail}}$ is at least 0.2. The tilted policy is used to collect rollouts and to compute stored action log probabilities and importance ratios. The critic retains the same task-value features and physical-cost targets.

Table 5: Shared PPO and PPO+Tilt training hyperparameters for the queueing experiment.

Hyperparameter	Value
Training steps per seed	200,000
Rollout length	1,000
Minibatch size	256
Actor / critic epochs per update	1 / 10
Initial actor / critic learning rates	3×10^{-4} / 10^{-3}
PPO clip range	0.2
Maximum gradient norm	0.5

Table 6: Columns report the margin, cumulative average task cost, tail occupancy, maximum queue length and the number of escapes over ten matched seeds. Scalar entries are seed medians; dashes indicate quantities unavailable for terminated runs.

Method	$\widehat{\mathcal{M}}$	Cost	P_{tail}	$q_{\max}$	Escapes
Vanilla AC	-0.3295	–	1.0000	–	10/10
PPO	-0.3216	8.0689	0.5403	6.54	0/10
PPO+Tilt	-0.0003	8.0608	0.5317	6.54	0/10
SVPO	0.0077	7.8891	0.2403	4.34	0/10

Vanilla AC reaches the physical escape threshold in all ten seeds. PPO (Schulman et al., 2017) completes training with a final worst-case finite-grid raw-policy margin of $\widehat{\mathcal{M}} = -0.3216$, consistent with its finite-horizon trajectories not reaching the far-tail region where this violation is measured. Applying Tilt raises the margin to -0.0003 while yielding a slightly lower task cost: the cumulative average costs are 8.0608 for PPO+Tilt and 8.0689 for PPO, consistent with Tilt's role as a minimal correction to an existing learner. SVPO maintains positive margins at every evaluated checkpoint in all ten seeds, with a final seed median of 0.0077, and achieves the lowest cost, tail occupancy, and maximum queue length. These results show that Tilt improves PPO's worst-case margin at similar task cost, while SVPO combines positive margins with improved task and tail metrics.